

University of Alberta

Mapping of the Region of Mouse Chromosome 6 Homologous to the
Human Cat Eye Syndrome Critical Region

by

Tim Footz



A thesis submitted to the Faculty of Graduate Studies and Research in partial
fulfillment of the requirements for the degree of Master of Science

in

Molecular Biology & Genetics

Department of Biological Sciences

Edmonton, Alberta

Fall, 1999.



National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-47030-X

Canada

University of Alberta

Library Release Form

Name of Author: Tim Footz

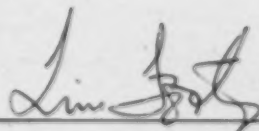
Title of Thesis: Mapping of the Region of Mouse Chromosome 6 Homologous to
the Human Cat Eye Syndrome Critical Region

Degree: Master of Science

Year this Degree Granted: 1999

Permission is hereby granted to the University of Alberta Library to reproduce
single copies of this thesis and to lend or sell such copies for private, scholarly or
scientific research purposes only.

The author reserves all other publication and other rights in association with the
copyright in the thesis, and except as herein before provided, neither the thesis
nor any substantial portion thereof may be printed or otherwise reproduced in
any material form whatever without the author's prior written permission.



128 Kirkwood Way

Edmonton, Alberta, Canada

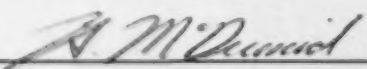
T6L-5P5

September 30, 1999.

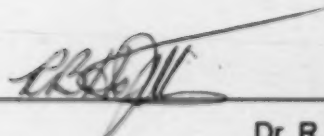
University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled *Mapping of the Region of Mouse Chromosome 6 Homologous to the Human Cat Eye Syndrome Critical Region* submitted by Tim Footz in partial fulfillment of the requirements for the degree of Master of Science in Molecular Biology & Genetics.



Dr. H. McDermid



Dr. R. Hodgetts



Dr. R. Godbout

Sept 30, 1999

Dedication

This manuscript is dedicated to my wife Stacey, who, apart from being a limitless source of encouragement, has thus far spent a significant portion of her life helping me to discover my own ambition and realize its potential.

Abstract

The smallest duplicated region of chromosome 22q11.2 responsible for the cat eye syndrome (CES) in humans is the proximal 2 Mb of the q arm (the "CESCR"). Coordinated strategies were undertaken for cloning genes in the CESCR and the homologous region of the mouse genome. Six new genes were placed on the physical map of the CESCR, including *BID*, whose homologue was mapped to mouse chromosome 6. Through hybridization and genomic sequence analysis approaches, the physical mapping of twelve mouse genes with homologues on human chromosome 22 aided the assembly of a mouse BAC contig covering >600 kb that contains the most distal gene in the CESCR. Homologous gene order and content are preserved along this contig, with disruption of conserved linkage beyond the gene *Il-17r*. The results suggest that an engineered mouse duplication will, to a large extent, provide a model for the smallest duplication that causes CES.

Acknowledgements

A number of individuals deserve recognition for their technical support of this project. Drs. Mary Barter and Lois Maltais (The Jackson Laboratory) aided in the genetic mapping of *Bid*. Dr. Bruce Birren (Center for Genome Research, Whitehead Institute/M.I.T.) was instrumental for production of the mouse physical map by donating both a set of membranes of the gridded BAC library and the BAC clones themselves. Dr. Bruce Roe (University of Oklahoma) and his technicians have provided the human and mouse genomic sequence of the CES and homologous regions. Drs. Rachel Wevrick and Mike Walter (University of Alberta) are acknowledged for lending the DNA and PCR resources needed for typing the Jax BSS backcross panel during the genetic mapping of *Bid*. A debt of thanks is owed to Angela Johnson, Dana Shkolny and Dr. M. Ali Riaz, researchers in the McDermid Lab who taught me all of the basic molecular protocols I use today, especially D.S. who performed the tissue culture steps during exon trapping. Fellow students Graham Banting and Polly Brinkman-Mills deserve special recognition for preparing certain molecular probes that were integral to assembling the mouse physical map. I also thank the Department of Biological Sciences for the Teaching Assistant scholarship that has financially supported me during this project. Finally, I extend my sincerest gratitude to Dr. Heather McDermid, my supervisor, for providing the most stimulating and cooperative work environment I have yet to experience, and for the integrity on which all of her scientific and professional endeavors are rooted.

Table of Contents

Chapter I: Introduction

Cat eye syndrome	1
CES: single-gene defect or contiguous gene disorder?	2
Duplications and dosage sensitivity	5
Mouse model systems	7
Physical mapping of the provisional CES critical region	9
Sequence annotation and comparative genomics	12
Research objectives	15

Chapter II: Materials and Methods

Exon trapping	20
Southern hybridization	20
DNA preparations	20
Sequencing	21
Dosage analysis	21
Expression studies	22
Cloning the extended 3' UTR of BID	22
Linkage mapping in mice	23
Cloning of Psl	23
Cloning of Ces38	24
Library hybridizations	24
Pulsed field gel electrophoresis	24
Computer analysis	24

Chapter III: Results

Cloning, sequencing and expression of BID, a gene near the distal boundary of the CESCO	29
Physical mapping of human BID outside of the CESCO	30
Genetic mapping of Bid on mouse chromosome 6	31

<i>The mouse homologues of three genes near the distal CESC boundary are linked on chromosome 6</i>	32
<i>Prediction and mapping of 2 novel genes and IL-17R from human genomic sequence (PAC 143I13)</i>	33
<i>Physical mapping of 14 mouse genes (in a region homologous to human 22q11.2) onto an assembled BAC contig</i>	35
<i>Prediction of a new gene (GAB/Gab) from human and mouse genomic sequence</i>	37
<i>Annotation of genes and gene duplications in the CESC and mouse chromosome 6</i>	42
 <u>Chapter IV: Discussion</u>	
<i>The CESC-homologous region in the mouse demonstrates conserved linkage of ten genes in the interval from IL-17r to Gab</i>	79
<i>Conservation of gene content between the CESC and mouse chromosome 6</i>	84
<i>Practical considerations for creating a mouse model of CES</i>	88
 Bibliography	96
Appendix I: GenBank Accession Numbers	109
Appendix II: Duplicated Segments within the Most Proximal 400 kb Sequenced from the CES Region	111

List of Tables

Table II-1. *DNA sequence of the PCR primers used in this study.* 28

Table III-1. *Calculation of BID dosage in the CES patient with the smallest duplication.* 57

Table III-2. *Calculation of relative BID 1.2 kb transcript levels in normal and CES patient cell lines.* 59

List of Figures

Figure I-1. <i>The CES chromosome</i>	17
Figure I-2. <i>Chromosome engineering protocol for producing a duplication of the proposed mouse genomic region sharing conserved linkage with the CESC.</i>	18-19
Figure II-1. <i>Exon amplification of CESC BAC clone KB70A6.</i>	27
Figure III-1. <i>BAC KB70A6 in relation to the genomic clones of the CESC being sequenced.</i>	49
Figure III-2. <i>Consensus sequence of the BID transcription unit(s).</i>	50-52
Figure III-3. <i>Southern hybridization of BID to chromosome 22.</i>	53
Figure III-4. <i>Northern analysis of human BID and mouse Bid.</i>	54-55
Figure III-5. <i>Dosage analysis of BID for the CES patient with the smallest duplication.</i>	56
Figure III-6. <i>Overexpression of BID in a typical CES patient with four copies of 22q11.2</i>	58
Figure III-7. <i>Sequence differences in the 3' UTRs of <i>Mus musculus</i> and <i>M. spretus</i> Bid.</i>	60
Figure III-8. <i>Mapping of Bid to mouse chromosome 6.</i>	61-62
Figure III-9. <i>Physical contig of BAC clones on mouse chromosome 6, in a region homologous to the human CESC on 22q11.2</i>	63
Figure III-10. <i>Partial genetic and cytogenetic maps of mouse chromosome 6.</i>	64
Figure III-11. <i>EST clusters on PAC 143I13.</i>	65
Figure III-12. <i>Comparative sequence analysis of the genomic regions surrounding human and mouse GAB/Gab.</i>	66-67
Figure III-13. <i>Northern analysis of GAB.</i>	68-69
Figure III-14. <i>Amino acid alignment of putative human and mouse GAB proteins.</i>	70
Figure III-15. <i>Comparative sequence analysis of the genomic regions surrounding human and mouse ATP6E/Atp6e.</i>	71-72

Figure III-16. <i>Comparative sequence analysis of the genomic regions surrounding human and mouse MTP/Mtp.</i>	73-74
Figure III-17. <i>Prediction of the genomic organization of BTPUTR.</i>	75-76
Figure III-18. <i>Organization of the genomic region of SAHL.</i>	77-78
Figure IV-1. <i>The physical order of genes within the CESCER and the homologous region (mCeser) in mouse.</i>	92-93
Figure IV-2. <i>The genetic map of mouse chromosome 6 demonstrating the conserved synteny with regions of the human genome.</i>	94-95

List of Gene Symbols

The names of many of the genes discussed in this study have not received approval of the human gene nomenclature committee. Instead, the monikers were derived from casual observations of the genes' properties. The standards for human and mouse gene notation were adopted, whereby all the letters in a human gene name are capitalized and italicized (e.g. *GENE*) but only the first letter is capitalized in the mouse homologue (e.g. *Gene*). Protein products follow the notation of all-caps but not italicized (e.g. *GENE*). These unpublished genes are as follows:

<i>SAHL</i>	(Like the Human homologue of the rat SA gene)
<i>BTPUTR/Btputr</i>	(Big Three Prime UnTranslated Region)
<i>PSL/Psi</i>	(Phosphatidyl Synthase Like)
<i>IDGFL/Idgfl</i>	(Insect Derived Growth Factor Like)
<i>CES38/Ces38</i>	(Cat Eye Syndrome gene for trapped exon # 38)
<i>CES11/Ces11</i>	(Cat Eye Syndrome gene for trapped exon # 11a)
<i>CTCO/Ctco</i>	(Cat eye syndrome Transcriptional COactivator)
<i>MTP/Mtp</i>	(Mitochondrial Transport Protein)
<i>GAB/Gab</i>	(Gene Abutting <i>BID</i>)
<i>KIAA0819/Kiaa0819</i>	(gene represented by the mRNA KIAA0819)
<i>Ng453</i>	(Novel Gene on mouse BAC 453L13)

List of Abbreviations

A - adenosine	DSCR - DS critical region
AGS - Alagille syndrome	dup - duplication
B - mouse strain C57BL/6Jei	ES cells - embryonic stem cells
BAC - bacterial artificial chromosome	EST - expressed sequence tag
bp - base pairs	F1 - first filial generation
BSA - bovine serum albumin	F2 - second filial generation
C - cytosine	G - guanosine
cDNA - complementary DNA	GAPD - glyceraldehyde-3-phosphate dehydrogenase
CES - cat eye syndrome	HAT - hypoxanthine aminopterin thymidine
CESCR - provisional CES critical region	HIV - human immunodeficiency virus
CF - cystic fibrosis	hr - hour
Chr. - chromosome	htgs - database of high-throughput genomic sequence
cM - centiMorgans	Jax BSS panel - The Jackson Laboratory's (BxS)F1 x S. backcross panel
CMT1A - Charcot-Marie-Tooth disease type 1A	kb - kilobase pairs
CNS - central nervous system	L - litre
dATP - deoxyadenosine trinucleotide phosphate	LCR22 - low copy-number repeat on chromosome 22
dbest - database of ESTs	LINE - long interspersed nuclear element
dCTP - deoxycytosine trinucleotide phosphate	LTR - long terminal repeat
del - deletion	M - unit of molarity
der(22) - derivative chromosome 22	Mb - megabase pairs
dic r(22) - dicentric ring chromosome 22	mCescr - region of mouse chromosome 6 homologous to
DGS - DiGeorge syndrome	
DNA - deoxyribonucleic acid	
DS - Down syndrome	

the CESCR
MDLS - Miller-Dieker Lissencephaly
syndrome
MIM - Mendelian Inheritance in Man
database
min - minutes
 μ g - micrograms
mg - milligrams
 $MgCl_2$ - magnesium chloride
MGD - mouse genome database
 μ l - microlitres
ml - millilitres
mM - millimolar
mol - moles
mRNA - messenger RNA
NCH - noncoding homologous
segment
NF1 - neurofibromatosis 1 gene
no. - number
nr - database of nonredundant
sequences
ONCR - renal-coloboma syndrome
ORF - open reading frame
PAC - P1-bacteriophage-based
artificial chromosome
PCR - polymerase chain reaction
PFGE - pulsed-field gel
electrophoresis
pmol - picomoles

RNA - ribonucleic acid
RT-PCR - reverse transcription PCR
s - seconds
S - mouse strain SPRET/Ei
SA - splice acceptor site
SD - splice donor site
SDS - sodium dodecyl sulphate
SINE - short interspersed nuclear
element
SSPE - sodium chloride sodium
phosphate + ethylene-
diaminetetraacetic acid
SSC - sodium chloride sodium citric
acid
SVAS - supra-avalvular aortic stenosis
T - thymidine
TAPVR - total anomalous pulmonary
venous return
TBS - Townes-Brocks syndrome
Tris-HCl - 2-Amino-2-
(hydroxymethyl)-1,3-
propanediol, hydrochloride
TOF - tetralogy of Fallot
URL - uniform resource locator
UTR - untranslated region
vol - volume
WWW - world wide web
YAC - yeast artificial chromosome

Chapter I: Introduction

Cat eye syndrome

Triplication of sequences from the proximal q arm of human chromosome 22 leads to the development of cat eye syndrome (CES; MIM115470). CES manifests as a rare association of relatively common birth defects. These defects include anal atresia (the lack of rupture of the anal membrane), ocular coloboma (the lack of closure of the optic fissure), preauricular (ear) pits and tags, congenital heart abnormalities (most commonly total anomalous pulmonary venous return [TAPVR] and tetralogy of Fallot [TOF]) and missing or underdeveloped kidneys (Schinzel *et al.*, 1981b). Other features in the CES spectrum include other urogenital defects, cleft palate, dysmorphic facial features such as downslanting palpebral fissures, wide-set eyes and low-set ears, and mild mental retardation. The comparison of CES patient phenotypes, even within families, reveals high variability in the severity of the symptoms. As well, many patients do not display the full range of defects such that a patient may present with only one or two of the diagnostic criteria (Schinzel *et al.*, 1981b). Despite the phenotypic variability, a common genetic mechanism appears to be the root cause of most cases of CES. The typical patient karyotype exhibits triplication of 22pter to 22q11.2 with the two extra copies inverted to comprise a bisatellited dicentric supernumerary chromosome (CES chromosome; Figure I-1; McDermid *et al.*, 1986). However, other genetic abnormalities can also lead to extra copies of this region and a CES-like phenotype, including supernumerary ring chromosomes (two extra copies, Mears *et al.*, 1995) and interstitial duplications (one extra copy; Knoll *et al.*, 1995; Reiss *et al.*, 1985; Lindsay *et al.*, 1995). In addition, children who inherit the der(22) chromosome as a result of 3:1 meiotic non-disjunction, from parents carrying the constitutional 11;22 translocation, show considerable phenotypic overlap with CES (Schinzel *et al.*, 1981a). The clinical features of these patients are characteristic facies (including deep-set eyes, flat nose and ear pits or tags), anal atresia, cleft palate, male genital

abnormalities and congenital heart defects, but no colobomata. These symptoms are due to possession of three copies of proximal 22q but also of distal 11q which may account for differences from the CES phenotype. Together, these findings suggest that overexpression of a dosage-sensitive gene or genes in the proximal q arm of chromosome 22 causes the features seen in CES.

Although the incidence of CES is rare (1 per 50-150,000 births; MIM115470) there are several reasons for concentrating efforts on identifying and characterizing the candidate genes. Chromosome 22 is likely to be the first human chromosome for which the complete sequence will be determined (H. McDermid, personal communication), thus allowing for sequence analysis to expedite the discovery of new genes in the CES region. It also offers the opportunity to study a disorder caused by a relatively small duplication, which may contribute to the understanding of more common conditions attributed to much larger duplications (such as trisomies or unbalanced translocations). More importantly however, the study of CES, a syndrome associated with multiple birth defects that are prevalent in their isolated forms, should lead to the uncovering of medically-important genes.

CES: single-gene defect or contiguous gene disorder?

Which gene products are responsible for producing the CES phenotype? Studying a condition caused by gene overexpression offers unique challenges to the standard positional cloning approach (i.e. initially pinpointing the area of the genome affected in a group of patients) to gene discovery. The majority of research on identifying haploinsufficient candidate disease genes involves sequencing patient DNA in hopes of finding genes interrupted or deleted by chromosomal rearrangements or affected by missense or nonsense mutations leading to decreased or inactive product (for examples, see Kishino *et al.*, 1997; Matsuura *et al.*, 1997; Kuslich *et al.*, 1999). However, the goal of CES research

is to determine which gene products, when overexpressed, have the potential to disrupt the developmental pathways of several embryonic tissues. The first step to address this "dosage-sensitivity" phenomenon might be to determine if overexpression of a single gene could account for all of the features of CES, or if the condition represents a "contiguous gene disorder" (Schmickel, 1986) whereby patients are aneusomic for a genomic segment containing more than one critical gene (reviewed in Budarf and Emanuel, 1997).

A typical example of a contiguous gene disorder is Miller-Dieker Lissencephaly syndrome (MDLS; MIM247200), characterized by specific facial features (bitemporal hollowing, prominent forehead, short nose, prominent upper lip and small jaw), polydactyly, heart and kidney defects and type I lissencephaly (absence of brain convolutions). MDLS is caused by microdeletions of 17p13.3. The associated lissencephaly trait is due to the inactivation of a single gene in this region, *LIS1*, in which point mutations cause isolated (non-syndromic) lissencephaly sequence (ILS; Lo Nigro *et al.*, 1997). The remaining features of MDLS are presumed to be caused by deletion of neighboring gene(s) on 17p13.3. Williams syndrome (WS) is another example of a multi-gene disorder, usually caused by a 2 Mb deletion of 7q11.23 (reviewed in Budarf and Emanuel, 1997). The *elastin* gene is within the commonly deleted region and is mutated in patients with isolated instances of the WS heart defect supravalvular aortic stenosis (SVAS; Li *et al.*, 1997). As well, patients with submicroscopic deletions involving only *elastin* and neighboring *LIM-kinase 1* (*LIMK1*) display only SVAS and the WS cognitive profile, without the typical facial features, infantile hypercalcemia or growth retardation associated with WS (Frangiskakis *et al.*, 1996). Since *LIMK1* is strongly expressed in the brain and mutations of *elastin* do not produce cognitive impairment, this strongly suggests that these two genes are each responsible for distinct features of a contiguous gene syndrome (WS) in the hemizygous state (Budarf and Emanuel, 1997).

Each of the main features of CES also occur in isolated, non-syndromic forms, and some have been attributed to genes in specific regions of the genome. For instance TAPVR, as studied in an extensive Utah kindred, displays

linkage to 4p13-q12 (Bleyl *et al.*, 1995; <http://www-medlib.med.utah.edu/reprogen/research/tapvr/index.html>), although the specific gene has not yet been identified. An unusual case is that of "isolated" tetralogy of Fallot, shown to be caused by a nonsense mutation in the gene *JAG1* (a ligand in the Notch signaling pathway) mapping to 20p12, or even by a deletion of the entire locus (Krantz *et al.*, 1999). Although these patients were ascertained because of congenital heart defects, further examination suggested they had mildly dysmorphic facial features similar to those of Alagille syndrome (AGS) patients, who normally also display specific liver, skeletal and other heart defects (MIM118450). Nonsense and deletion mutations in *JAG1* also cause typical AGS, thus suggesting that other patients with apparently isolated heart defects may not be diagnosed with AGS due to the extreme phenotypic variability of the syndrome. Studying CES could reveal a similar situation, whereby the full spectrum of features is caused by a single gene with incomplete penetrance and variable expressivity. However, it is also possible that separate genes could each be responsible for producing a specific trait or subset of traits, and like the effects seen by mutations of *JAG1*, phenotypic variability would result from the duplication of specific genes.

Mounting evidence indicates that single gene defects can be the cause of syndromes with effects on multiple tissues. Two syndromes with considerable phenotypic overlap with CES have been shown to be due to mutations in single genes. Townes-Brocks syndrome (TBS; MIM107480) is also characterized by anal atresia and renal defects, in addition to hand, foot and ear anomalies. Kohlhase *et al.* (1998) demonstrated mutations in the putative transcription factor *SALL1* in TBS patients. Another transcription factor, *PAX2*, is mutated in patients with Renal-Coloboma syndrome (ONCR; MIM120330) who show renal defects with optic nerve colobomata (sometimes seen in CES patients). Interestingly, mutations in mouse *Jag1* result in ocular coloboma, which has not been observed in AGS patients, as well as vascular defects (Xue *et al.*, 1999). This wealth of information from mutational studies suggests that it is possible for CES to result from the altered expression of a single gene, perhaps one that

interacts (directly or indirectly) with the genes responsible for the above conditions. Candidates could include components of signaling pathways, such as transcription factors, extracellular receptors or secreted ligands. However, it is also possible that multiple genes are responsible for producing all of the CES features.

Duplications and dosage sensitivity

Few medical conditions compatible with extended postnatal life arise from recurrent gene duplications. Two of the most intensively-studied are Charcot-Marie-Tooth Disease type 1A (CMT1A; MIM118220), a demyelinating neuropathy resulting from duplication of the *PMP22* gene, and Down syndrome (MIM190685), caused by complete or partial trisomy of chromosome 21.

Peripheral myelin protein 22 (*PMP22*) is an integral membrane protein of the myelin of peripheral nervous system Schwann cells (MIM601097). Mutations in this gene, causing different forms of inherited neuropathies, suggest that it must be maintained at diploid dosage for normal motor nerve conductance. It was first considered a candidate for CMT1A after noticing that a similar mouse mutant phenotype, the "Trembler" neuropathy, might be caused by mutation in a homologous mouse locus (Vance, 1991). In support of this, the Trembler mutation was known to map to chromosome 11 in a region sharing conserved linkage with human chromosome 17p11.2, which shows linkage to CMT1A (Vance *et al.*, 1991). Gene duplication was later implicated as the mutational event leading to CMT1A by Patel *et al.* (1992), although patients without a detectable duplication have been identified (reviewed in Suter and Patel, 1994). *PMP22* lies in a 1.5 Mb interval flanked by repeated elements (CMT1A-REP) involved in the unequal crossing-over that leads to reciprocal duplication (CMT1A) or deletion (hereditary neuropathy with liability to pressure palsies) of the gene. The dosage-sensitive nature of this gene was clearly shown by pronuclear injection of human *PMP22*-containing YAC DNA to create mice with a

mutant phenotype closely mimicking CMT1A (Huxley *et al.*, 1996). Missense mutations within the coding region of *PMP22* can also be found in some CMT1A patients (Valentijn *et al.*, 1992; Roa *et al.*, 1993), suggesting that an increased level of transcription is not the only means of replicating the defects usually caused by duplication of *PMP22*. It is therefore possible that CES patients without a cytogenetically-visible duplication (Franklin and Parslow, 1972) may possess point mutations in a single critical gene.

Trisomy 21 results in Down syndrome (DS), the most frequent cause of mental retardation. Other features of DS include heart defects, characteristic craniofacial anomalies, skeletal defects, susceptibility to leukaemias and "premature ageing". The q arm of chromosome 21 is ~37 Mb long and could contain between 700 and 1000 genes (reviewed in Antonarakis, 1998) presenting an onerous task for identifying DS candidate genes. The search for the dosage-sensitive genes on 21q has therefore relied on cases of partial trisomy 21 which have delineated DS critical regions (DSCR) for subsets of defects to areas as small as 4 Mb. The smallest DSCR contains genes which contribute to mental retardation and several of the facial and skeletal abnormalities. The involvement of these genes in contributing to the DS phenotype is suggested by the production of transgenic mice overexpressing single DSCR genes. Mice carrying a human cDNA-based *ETS2* transgene (encoding a proto-oncogenic transcription factor expressed highly in newly forming cartilage) develop craniofacial abnormalities homologous to those seen in DS patients (reviewed in Kola and Hertzog, 1997). As well, the homologue of the *Drosophila* protein kinase *minibrain* (*MNBH/DYRK1*) was implicated as a dosage-sensitive gene responsible for learning and memory defects, by mice carrying a fragmented yeast artificial chromosome (Smith and Rubin, 1997).

It is clear that mouse transgenic models are an invaluable resource for dissecting the pathophysiology of human duplication disorders. While the identification of *de novo* mutations in a single candidate gene, altering conserved residues or causing premature truncation or decreased expression of the protein product, is often enough evidence to conclude a gene's involvement in human

disease, the lack of such informative patients necessitates using model organisms to study alterations of the candidate genes. The large-scale molecular defect associated with CES patients therefore lends itself to be studied in a mouse model system.

Mouse model systems

The advantages of making murine models for human disease are numerous. Firstly, the mouse represents the best studied example of mammalian development, in terms of mutant phenotypes, gene defects and gene interactions. As numerous mouse models have demonstrated (e.g. Waardenburg syndrome/Spotch, cystic fibrosis, Alzheimer's, diabetes, atherosclerosis and Hirschsprung disease; reviewed in Erickson, 1996; Bedell *et al.*, 1997), humans and mice have generally preserved many developmental pathways and gene functions over the ~80 million years since the divergence of the species. This allows mutations in orthologous (functionally homologous) genes to produce homologous mutant phenotypes. These mutant phenotypes can then be exploited in an experimental system for the purpose of designing therapeutic strategies.

Secondly, stable integration of recombinant DNA into the genome of mice has become a relatively straightforward way of testing a plasmid- or YAC-borne transgene in a mammalian model organism (reviewed in Peterson *et al.*, 1997). The technique of microinjection of highly purified DNA into mouse male pronuclei can allow the transgenesis of mice carrying up to 2 Mb of human DNA (in theory). This method is suitable not only for producing homologous mutant phenotypes of overexpression disorders, it is also practical as a tool to confirm the orthologous relationship of human and mouse homologues by transgenic rescue, thus strengthening the applicability of proposed therapies (see Manson *et al.*, 1997 for a description of *Cftr*-deficient mice partially rescued by the homologous human transgene).

Another useful method for modeling human genetic defects is site-directed manipulation of mouse embryonic stem (ES) cells *in vivo*, through homologous recombination (reviewed in Rossant and Nagy, 1995). ES cells can be maintained in culture and subjected to integration, and subsequent antibiotic-selection, of plasmid DNA into homologous sites of the genome. The cells can then later be incorporated into blastocysts to produce chimaeric mice, with transmission of the altered genome into germline stem cells, to be propagated by controlled mating. While this "basic" gene-targeting method is efficient for creating base substitutions and small deletions and insertions, a modification of the design allows for the production of large deletions and duplications up to 3-4 cM (Ramírez-Solis *et al.*, 1995) as well as translocations (Smith *et al.*, 1995). "Chromosome engineering" (Figure 1-2) introduces recombinogenic lox P sequences with adjacent selectable markers to sites flanking the region to be reciprocally deleted/duplicated, by sequential ES cell targeting events (Ramírez-Solis *et al.*, 1995; Rossant and Nagy, 1995). Integration of each transgene into a separate homologous chromosome (i.e. in *trans*) will allow for the creation of genetically-balanced products capable of stable mitotic divisions. Transient expression of transfected Cre recombinase mediates intrachromosomal recombination between the lox P sites, thereby creating two derivative chromosomes, one carrying a deletion of the region, the other a duplication. Segregation of these chromosomes is then monitored by molecular methods upon generation and mating of chimaeric mice. This procedure depends on physically mapping the homologous mouse region of interest, and would benefit from knowledge of gene order at the extremities of the interval for correct construction of the lox P cassettes. Although more technically-demanding than plasmid or YAC transgenesis for modeling a duplication, chromosome engineering may produce the mutant phenotype more faithfully because it does not rely on assumptions of correct expression and proper functioning of human transgenes in murine cells. It does rely however on the duplicated mouse genes acting in a manner similar to the human homologues. While pronuclear injection can allow for numerous copies of the transgene to integrate into the mouse

genome and conceivably produce a lethal effect or, alternatively, (co)suppression of the locus (Garrrick *et al.*, 1998; Bingham, 1997), the precision of chromosome engineering to control the copy number of the manipulated region is monitored by drug selection in culture and Mendelian segregation in mated animals.

A physical map identifying all of the genes in the CES region will be crucial to the selection of the most promising candidates on which to focus research, so as to minimize the cost and effort put into developing a mouse model. As of early 1998, prior to this study, the only gene mapped to the region found on all CES chromosomes was *ATP6E* (Baud *et al.*, 1994) which encodes a subunit of the vacuolar proton pump (H^+ -ATPase), involved in acidification of intracellular compartments and bone resorption by osteoclasts. Its ubiquitous expression suggests it is unlikely to contribute to the tissue-specific defects of CES when overexpressed. As well, intravenous drug-induced disruption of the H^+ -ATPase complex in rats leads to inhibition of bone resorption (modeling osteoporosis), disturbed locomotor control and cyanosis and convulsions upon increasing doses (Keeling *et al.*, 1997). The lack of such phenotypes in CES patients suggests a lack of involvement from overexpression of *ATP6E*. Research and development of a mouse model of CES will benefit from the construction of a more detailed physical map including all of the transcribed sequences of 22q11.2. As well, physical mapping of the mouse homologues may reveal conserved linkage of these genes entailing chromosome engineering as the way to most closely model a CES duplication.

Physical mapping of the provisional CES critical region

An international consortium has been organized to achieve the goal of sequencing the entirety of chromosome 22 (<http://www.sanger.ac.uk/HGP/Chr22>). As the preparation of sequence-ready DNA clones depends on localization of genetic and physical markers to maps of overlapping chromosome fragments, high-density YAC contigs were established (Collins *et al.*, 1995; Bell

et al., 1995) and provide frameworks for even higher-resolution maps of smaller-insert clones which are more amenable to sequencing. Portions of the chromosome have been allocated to four genome sequencing centres, including the University of Oklahoma which is responsible for much of proximal 22q (from D22S50 to GNAZ-BCR; <http://www.genome.ou.edu/maps/ch22.html>). This group is currently sequencing clones (provided by the McDermid lab) that should contain the genes responsible for CES (<http://www.genome.ou.edu/maps/ces.html>).

The minimal duplicated region of 22q11.2 causing CES was defined in patient 25105 who possessed an unusual supernumerary double ring chromosome (Mears *et al.*, 1995). Molecular studies of the size of the duplication outlined a ~2 Mb region from the centromere to marker D22S57 (Mears *et al.*, 1995; McDermid *et al.*, 1996). Since the patient's phenotype included all of the cardinal features of CES, including coloboma, anal atresia, ear tags and pits, cardiac and urogenital defects and cleft palate, this region should contain all or most of the genes responsible for the phenotypes observed in patients with the inverted supernumerary chromosome. This critical region can be refined further by patient S.K., who carries an interstitial duplication with the proximal boundary located ~1 Mb from the centromere (Knoll *et al.*, 1995), distal to marker D22S795 (Mears *et al.*, in preparation). Although S.K.'s phenotype did not include coloboma or anal atresia, he did present with TAPVR, preauricular pits, absent right kidney, mental retardation and CES-like facial malformations (Knoll *et al.*, 1995). Therefore, the region from D22S795 to D22S57 (an ~1 Mb interval) can be accepted as a "provisional" CES critical region (CESCR) containing genes that produce at least some of the features of CES when duplicated. This interval may encompass all of the genes responsible for the full spectrum of CES phenotypes owing to the phenotypic variability and reduced penetrance of the associated traits. As well, the extent of S.K.'s duplication into band q12 (Knoll *et al.*, 1995) may have somehow suppressed the mutant phenotype. Alternatively, it is possible that in patient S.K., duplicated genes distal to the CESCR may have contributed to some of his features.

It is also likely that the most proximal 1 Mb of 22q is extremely poor in functional genes. The pericentromeric regions of human chromosomes are rich with repetitive satellite DNA and duplicated gene segments which may contribute to specifying the operative location of the centromere (reviewed in Eichler, 1999). The duplicated segments studied thus far appear to contain genomic regions complete with intervening sequences, unlike transposed processed pseudogenes. For example, six of nine *NF1*-related loci throughout the genome are located in pericentromeric regions, contain only a medial portion of the gene and are rife with point mutations likely rendering them pseudogenes (Régnier *et al.*, 1997). As well, Eichler *et al.* (1996) described a duplication of 27 kb shared between Xq28 and 16p11.1, containing exon and intron sequences from two genes. The regions are 94.6 % identical, suggesting the duplication occurred 7-10 million years ago, and the 16p11.1 paralogues have been rendered pseudogenes, implicating Xq28 as the ancestral locus (Eichler *et al.*, 1996). Accumulation of gene duplications in the pericentromeric region of 22q (e.g. *von Willebrand factor* [Mancuso *et al.*, 1991] and *NF1* [Régnier *et al.*, 1997]) may have also occurred by similar mechanisms causing the region to be composed of numerous nonfunctional sequences. Since such gene transpositions are recent, after the divergence of mouse and man, this implies that any functional genes created in this manner in proximal 22q will not have orthologues in mice (by definition), thus hindering their experimentation as CES candidates.

The first steps in developing a transcript map of the CESCO were construction of a framework YAC map (McDermid *et al.*, 1996) followed by the assembly of a sequence-ready BAC and PAC contig across the region (see Figure III-1; Johnson *et al.*, 1999). While this work and subsequent sequencing were underway (at the University of Oklahoma), identification of transcribed sequences was also initiated through exon amplification. This method, which identifies sequences from cloned genomic inserts that contain functional splicing signals, yielded the discovery of several genes and pseudogenes within the CESCO (Riazi, 1998; Brinkman-Mills, 1999), including a putative growth factor with significant homology to adenosine deaminases (Riazi *et al.*, submitted). A

gene, *CTCO*, that shows amino acid sequence similarity to transcriptional co-activators, such as *CREB binding protein* (MIM600140) and *histone acetyltransferase 1* (MIM 603053), was located within partial sequence generated from BAC clones mapping to the CESCO (G. Banting, unpublished). As the identification of putative candidates for CES progresses, their ability to contribute to pathogenesis of CES-like features in mice must be tested. Dosage-sensitive correlations of the candidates to mutant phenotypes, as well as the characterization of novel genes in the CESCO, may be aided by physical mapping of, and comparison to, their mouse homologues.

Sequence annotation and comparative genomics

One of the goals of the Human Genome Project is the complete sequencing and characterization of all human genes (<http://www.nhgri.nih.gov/HGP>), in a manner similar to the coordinated effort to study chromosome 22. The quickest way to actualize the list of transcribed sequences from the CESCO is to determine the similarity of the genomic DNA to known genes in humans and other species and to randomly-selected cDNA clones. Such sequence analyses are quickly superceding other methods in the ease of identifying transcribed regions of the genome, such as exon amplification or hybridization of cDNAs to the region of interest. Sequence similarity searches can be routinely performed by BLAST analysis (Altschul *et al.*, 1990) with computer databases of genes, genomic clones and expressed sequence tags (ESTs) from the 5' and 3' ends of random cDNAs. Selected BAC and PAC clones mapping to the CESCO are currently being sequenced by Dr. B. Roe at the University of Oklahoma (http://www.genome.ou.edu/hum_totals.html) as an integral part of the "positional cloning" approach to identify CES candidate genes (positional cloning and database searching are reviewed in Rastan and Beeley, 1997). Genes previously identified in the CESCO, as well as those that emerge from sequence annotation, will be further characterized by gene prediction and cross-species

comparisons ("comparative genomics"), which rely on high-quality genomic sequence. Submission of genomic sequence to a battery of online gene/exon prediction programs (reviewed in Claverie, 1997) along with identification of EST hits (<http://www.ncbi.nlm.nih.gov/dbEST/index.html>) can be very successful for discovering genes through an "*in silico*" approach. Prediction programs such as GENSCAN (Burge and Karlin, 1997) and MZEF (Zhang, 1997) are very efficient at identifying coding exons bounded by proper splicing signals, based on the likelihood that the nucleotides are involved in nonrandom combinations. However, a thorough description of a gene's structure and expression profile requires "hands-on" molecular research and is greatly enhanced by comparative analysis to the genomic sequence of model organisms (Carver and Stubbs, 1997; Clark, 1999), which can be further manipulated for mutational analysis and reporter gene assays.

The value of comparative genomics in gene characterization was most elegantly demonstrated by Oeltjen *et al.* (1997) who sequenced ~90 kb from human and mouse chromosomes X around the *BTK/Btk* locus to confirm the conservation of promoter, exon and intron sequences from five adjacent functionally unrelated genes. The "weighted" measure of conservation was calculated as the average percent identity relative to the combined length of the noticeably homologous fragments. This measure for non-coding regions resulted in 73 % identity (for the total of stretches >50 bp each showing at least 60 % identity), a surprisingly high level, considering the measure for coding sequences showed only 87 % identity (Oeltjen *et al.*, 1997). These findings suggest that a substantial amount of functional information is not confined to coding exons. In all, 16% of the locus is comprised of homologous non-coding sequence, and would not have been identified by gene prediction or sequence comparison to cDNAs. Hardison *et al.* (1997) pointed out that, "clusters of invariant or slowly changing positions in the aligned sequences are 'phylogenetic footprints', which are reliable guides to important regulatory regions". In corroboration of this proposal, Oeltjen *et al.* (1997) demonstrated regulatory activity in the first intron of *BTK*. This 2.5 kb fragment was tested in a luciferase reporter assay in human

cell lines and shown to contribute to lineage-specific downregulation of *BTK*. Other recent studies have reported on the conservation of non-coding homologous sequences from mouse and human (Ansari-Lari *et al.*, 1998; Jang *et al.*, 1999), serving to remind us of the complexity of gene expression and suggesting additional targets (apart from exons and splice junctions) with which to search for disease-causing mutations.

To what extent is the organization within a large-scale genomic region conserved between mouse and man? The realities of conservation of genetic linkage amongst mammals have been studied intensively in this recent era of "genomics" (McKusick, 1997; Carver and Stubbs, 1997). Regions up to 11-13 Mb (Oakey *et al.*, 1992) have been shown, by long-range restriction mapping, to preserve homologous gene content, order and spacing since rodent-primate divergence. However, rearrangements within blocks of conserved synteny have in some cases led to local differences. For example, gene order in the human 22q11 deletion (DiGeorge) syndrome region, just distal to the CESC, and the homologous region on mouse chromosome 16, has undergone several inversions throughout the separate mammalian evolutionary lineages (Puech *et al.*, 1997). As well, differences in gene content in this region are suggested by the lack of identification of the mouse homologue of *CLTD* (based on cross-species hybridization; Puech *et al.*, 1997) and by significant sequence differences in the *TSK1/Tsk1* and *DGS-H* loci (Galili *et al.*, 1997). Whether the mouse homologues of the genes within the CESC are themselves linked, awaits completion of the human transcript map to aid the identification of interspecific probes, and subsequent physical mapping in the mouse (and also sequencing of both species). It is hoped that there will be a single region encompassing the CESC-homologues as this will simplify a chromosome engineering-based approach for creation of a mouse model for CES (Ramírez-Solis, *et al.*, 1995).

Research objectives

The main goal of this project is to determine the feasibility of modeling CES through an engineered duplication in the mouse. This process involves the following steps:

- **Identifying genes flanking the human CESC, by exon trapping and genomic sequence analysis.** From these genes, probes will be identified that hybridize to mouse DNA in order to build contigs of physical clones that contain CESC-homologous sequences. Should the mouse genome contain a single region of conserved linkage (with respect to the CESC), flanking sequences will be identified in the mouse to facilitate construction of lox P cassettes that could be used to engineer a duplication with the Cre/lox P system.
- **Determining the extent of conservation of linkage between CESC genes and their mouse homologues.** Concurrent gene discovery in human and mouse will reveal if each gene has homologous counterparts in both species. If the mouse homologues are as closely linked as the CESC genes, each region will be examined for preservation of gene content (i.e. promoter, coding and regulatory DNA sequences), and for any significant genomic rearrangements that may have disturbed the gene order or orientation. The construction of a physical map of the CESC-homologous region with mouse Bacterial Artificial Chromosomes (BACs) will aid the fine mapping of genes, and produce a contig of sequence-ready clones. If the homologous genes exhibit a high level of conservation, targeted duplication of the mouse genome should mimic the production of CES.

- **Characterizing newly-discovered genes by sequence analysis.**

Computer-driven searches for homology of human and mouse genomic sequence to known genes or transcribed sequences will help construct a complete transcript map of the CESC and the homologous mouse region. Examination of human and mouse genomic sequence will allow for comparative analysis to refine gene structure determinations by ascertaining slowly-evolving (and therefore functional) elements. The likelihood of the involvement of candidate genes in contributing to the CES phenotype may be suggested by identifying similarity to known disease-causing genes.

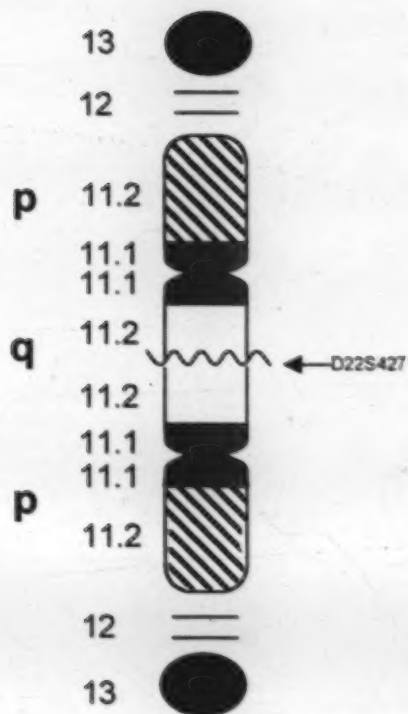
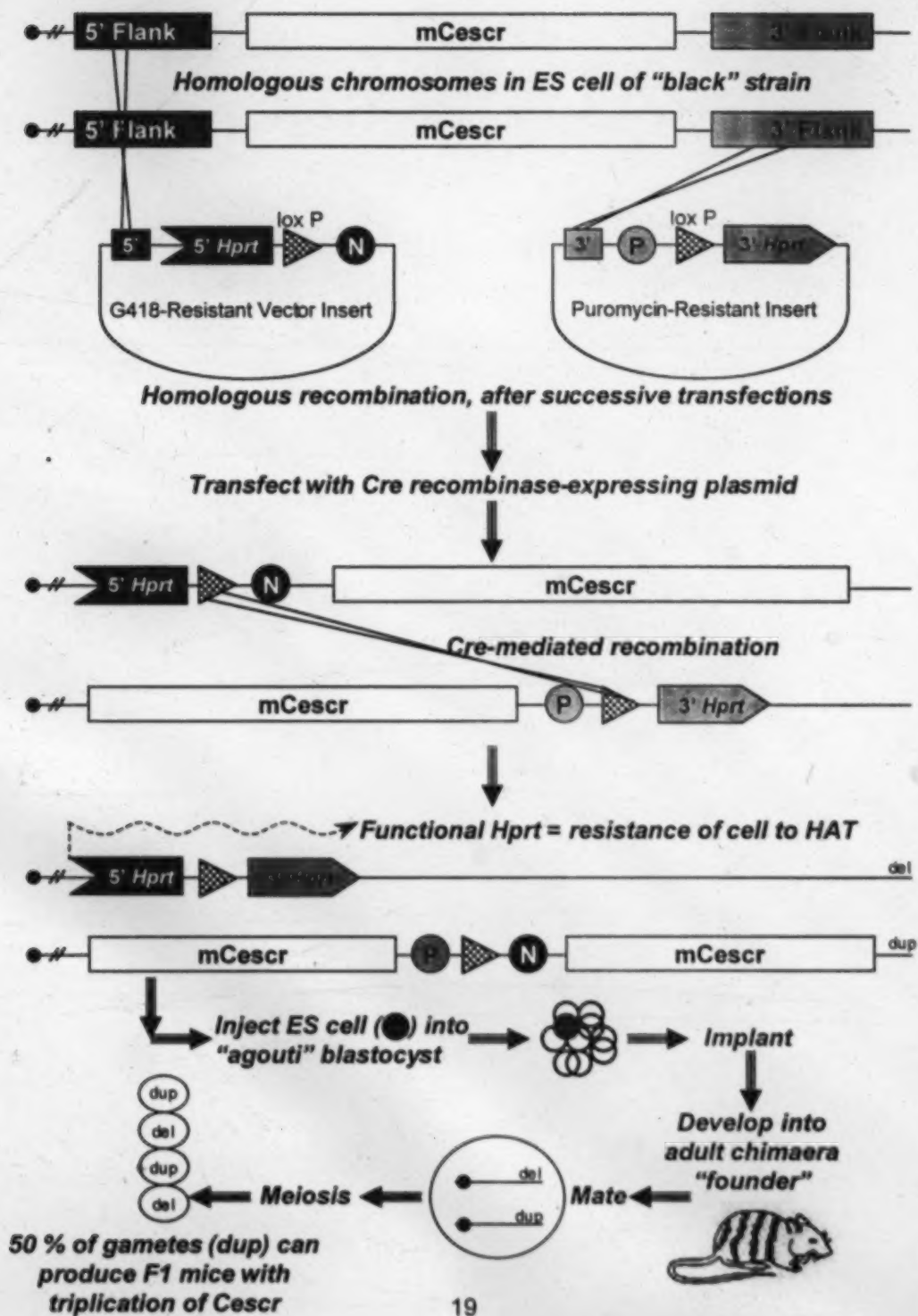


Figure I-1. The CES chromosome. Patients carrying the supernumerary chromosome depicted above will possess four copies of the region from the tip of the p arm, to locus D22S427 on the q arm.

Figure I-2. Chromosome engineering protocol for producing a duplication of the proposed mouse genomic region sharing conserved linkage with the CESCO. Sequential integration of the half-*Hprt*/lox P gene cassettes into the genome of ES cells from a "black" mouse strain is shown occurring in *trans* (i.e. each cassette integrates into a separate homologous chromosome). The 5' integrant is selected for by resistance to G418, conferred by the neomycin-resistance gene (N). The 3' integrant is selected for by resistance to puromycin (P). Cre-directed recombination between the lox P sites allows reconstruction of the functional *Hprt* gene on the deleted allele (lox P is spliced out post-transcriptionally) and therefore resistance of the cell to HAT selection in culture. Fortuitous incorporation of recombinant "black" ES cells into "agouti" strain blastocysts will contribute to formation of the chimaera's germline so that the recombinant chromosomes may comprise part of the gamete pool. Segregation is shown occurring in a recombinant meiocyte resulting in 50 % of the meiotic products carrying the duplication of the CESCO-homologous region (mCescr). Fertilization by male founders leads to some F1 progeny with three copies of the Cescr, which can then be intercrossed, if desired (25 % of the possible F2 progeny [F1 dup x F1 dup] would contain four copies of the mCescr, similar to the triplication often seen in CES patients). This protocol was adapted from Ramírez-Solis *et al.* (1995).



Chapter II: Materials and Methods

Exon trapping

BAC KB70A6 was subcloned with *Sst*I or *Bam*HI into the intron cloning site downstream from the SV40 promoter of the pSPL3B vector (Figure II-1; Burn *et al.*, 1995) and electroporated into COS-1 cells. RT-PCR on total RNA from transfected cells was performed according to the protocols supplied with the Exon Trapping System (Life Technologies, Inc.). Insert-containing products (i.e. >177 bp) were hybridized back to KB70A6, to confirm their origin, before sequencing.

Southern hybridization

Hybridization of exon and cDNA probes to total human or bacterial clone DNA was carried out in a manner identical to the procedures described in Mears *et al.* (1994) with the following exceptions: 1) DNA probes were purified in 0.7 or 1.0% low melt agarose (SeaPlaque, FMC); 2) transferred DNA was separated on 0.7 - 2.0% agarose; 3) the first post-hybridization wash consisted of 1.5 x SSC / 0.2% SDS; and 4) all Southern blot washes ranged from 5 - 20 minutes. Human:human or mouse:mouse hybridizations were performed at 65°C whereas interspecific hybridizations were done at 50-55°C with washes no higher than 60°C.

DNA preparations

Total genomic DNA was provided by other researchers in the McDermid lab and was isolated by standard means (Riazi, 1998). Plasmid DNA was isolated using standard methods (Sambrook *et al.*, 1989) or with the aid of

QIAprep Spin Miniprep columns (QIAGEN). Mouse BAC DNA was isolated with a modification of the protocol provided with QIAGEN TIP-500 columns as follows: 100 ml each of solutions P1, P2 and P3 were applied to cells from a 500 ml culture; DNA was then precipitated with 0.1x vol of 3M sodium acetate (pH 7.0) and 0.7x vol isopropanol and resuspended in 2 ml TE (pH 8.0) plus 10 ml QBT buffer; purification through the column followed by elution with 15 ml QF buffer at 65°C; this was followed by another precipitation (see above) and resuspension in 500 µl distilled water plus 0.1 mg RNase A at 37°C for 1 hr.

Sequencing

The sequence of the *BID* cDNA and exon-containing pAMP10 clones was determined using the ThermoSequenase kit (Amersham, Inc.) involving gene-specific and vector-based primers, separated on 8% polyacrylamide-urea gels and visualized on Biomax film (Eastman Kodak). Sequence for every other cDNA was determined by other researchers at the University of Alberta with Thermosequenase Kit RPN2438 (Amersham, Inc.) and analyzed on a Licor Sequencer model 4200LD.

Dosage analysis

Human genomic DNA from lymphoblastoid cell lines was digested with *HindIII*. Southern hybridization was carried out by probing simultaneously with a 0.22 kb *SalI/BglII* fragment of the pAMP10-based clone of *BID*'s 211 bp exon and the nonsyntenic reference probe D21S15 from chromosome 21 (Stewart *et al.*, 1985). Densitometric analysis of the signals was performed using pixel volumes/intensities obtained from a Molecular Dynamics PhosphorImager 445SI and ImageQuant v1.1 for Macintosh. A consistent field of measurement entirely within each band of a phosphorimage (in the form of a rectangle) was used to

calculate pixel values of the scanned image. Data sets were compared by a statistical method described in Mears *et al.* (1995).

Expression studies

Poly A⁺-containing human multiple tissue Northern blots (Clontech, Inc.) were hybridized with radiolabelled *BID* cDNA according to the manufacturer's directions. The human *BID* cDNA was also hybridized to the human lymphoblastoid Northern blot. Isolation of total RNA from mouse tissues (strain CD1) and human lymphoblast cell lines was performed using TRIzol Reagent (Life Technologies, Inc.). Immobilization to GeneScreen Plus membranes (DuPont) followed a described procedure (Ausubel *et al.*, 1989). The mouse cDNA for *Bid* was hybridized to the mouse tissue Northern blots. Human cDNAs for *GAB* were radiolabeled with ³²P-dATP and hybridized to human multiple tissue Northern blots (Clontech, Inc.) according to the STRIP-EZ DNA kit's directions (Ambion) to allow for easy cleavage, at modified dCTP residues, and removal of the probes.

Each hybridization was performed at 42°C in a solution consisting of 50% formamide, 5x SSPE, 10x Denhardt's, 2% SDS and 80 mg/L herring sperm DNA. The low stringency wash buffer was 2 x SSC/1% SDS. The high stringency wash buffer was 0.1x SSPE/0.1% SDS. See the legends for Figures III-4 and III-13 for reaction, wash and X-ray exposure times.

Cloning the extended 3' UTR of BID

A 452 bp product was amplified from BAC KB70A6, total human DNA and DNA from a human chromosome 22/hamster hybrid cell line (data not shown) with the primers BIDENDF and BIDENDR (Table II-1), which were designed from the 3' EST for cDNA clone 503784. The primers were added at 50 pmol each to

cycling conditions of: 30 cycles of (94°C for 30 s, 61°C for 30 s and 72°C for 1 min), followed by 10 min at 72°C.

Linkage mapping in mice

The (C57BL/6JEi x SPRET/Ei)F1 x SPRET/Ei ("Jax BSS") backcross panel created and distributed by The Jackson Laboratory (Bar Harbor, ME) was employed to map *Bid* by *Dde* I restriction analysis of PCR products. Overlapping PCR products from the 3' UTR of *Bid* were amplified and sequenced from *Mus musculus* (C57BL) and *M. spretus* (SPRET) to identify restriction fragment length variants (see Figure III-7). The primers MBID2F and MBID3R2 (Table II-1) amplify an ~779 bp product from the 3' UTR of *Bid*, which contains a nucleotide difference between the two species in a *Dde* I recognition site in *M. spretus* that is altered in *M. musculus*. PCR conditions consisted of 20 µl reactions containing 62.5 µg DNA, 25 pmol of each primer, 0.5 µl Taq polymerase, 45 mM Tris-HCl, pH 9.0, 1.35 mM MgCl₂, 0.36 mM β-mercaptoethanol, and 1.8 µg BSA. The reactions were performed in M.J. Research, Inc. thermal cyclers at 94°C for 5 min plus 30 cycles of [94°C for 45 s, 55°C for 30 s, and 72°C for 1 min], followed by 72°C for 10 min. Haplotypes were analyzed and map position determined by The Jackson Laboratory.

Cloning of *Psl*

Amplification of a 165 bp product corresponding to a part of the coding region of *Psl* was achieved with a mouse genomic DNA template. The primers MYOF and MYOR (Table II-1), designed from the sequence of EST 1050354-5' (for which the corresponding cDNA clone was not available), were added at 167 pmol per primer to the following reaction conditions: 30 cycles of [95°C for 1 min, 65°C for 30 s and 72°C for 30 s], followed by 72°C for 10 min.

Cloning of Ces38

A 1.6 kb product was amplified from mouse BACs 541L22 and 509P19 with 50 pmol each of the primers M38F and M38R (Table II-1), designed from partial sequence of 541L22 that is homologous to the human PAC 238M15. A "touchdown" PCR protocol was implemented with 10 cycles of [94°C for 1 min, 68°C (minus 0.5°C/cycle) for 30 s and 72°C for 90 s] followed by 25 cycles of [94°C for 1 min, 61°C for 30 s and 72°C for 90 s] and 72°C for an additional 10 min.

Library hybridizations

A mouse BAC library for strain 129SV (Research Genetics, Inc.) was kindly donated by Dr. B. Birren. Hybridizations to eight high-density membranes (containing >200,000 double-spotted unique clones in total) were performed in 5x SSC / 5x Denhardt's / 0.5% SDS with washes as described for Southern analysis above. Positive clones were also provided by Dr. B. Birren.

Pulsed field gel electrophoresis

Separation of BAC insert DNA on 0.8 or 1.0% agarose gels followed the conditions described in Johnson *et al.* (1999), or with a 3-30 s pulse for a 20-23 hr run time.

Computer analysis

ZAPMAP for Macintosh (B. Kraus, unpublished) was used for graphical assembly of the human and mouse physical contigs.

GeneTool v1.0 for Macintosh was employed for sequence chromatograph interpretation, *in silico* gene modeling and the color-based annotation of genomic sequence used in some figures.

Human and rodent interspersed repeats were detected and masked using RepeatMasker (A.F.A. Smit & P. Green, unpublished) at the WWW interface <http://ftp.genome.washington.edu/cgi-bin/RepeatMasker>. Simple sequence repeats and low complexity regions were not masked.

The Basic Local Alignment Search Tool (BLAST) algorithm (Altschul *et. al*, 1990) was accessed at <http://www.ncbi.nlm.nih.gov/blast/blast.cgi?Jform=1> for DNA and protein similarity searches against the non-redundant (nr), expressed sequence tag (dbEST) and high throughput genomic sequence (htgs) databases using the default parameters (or without "filtering"). The "BLAST 2 Sequences" tool (Tatusova and Madden, 1999) for pairwise alignments was accessed at <http://www.ncbi.nlm.nih.gov/gorf/bl2.html> using the default parameters (or without "filtering"). All sequences were repeat-masked before querying a database. At least one sequence was repeat-masked when compared against another by "BLAST 2 Sequences".

ALIGN, accessed at <http://vega.cr.bm.cnrs-mop.fr/bin/align-guess.cgi>, was used to calculate percent identity between compared amino acid or DNA sequences.

GeneDoc v2.1.0 for Windows (<http://www.cris.com/~Ketchup/genedoc.shtml>) was employed for representation of the *BID* cDNA and GAB protein alignments, arranged by the MAP WWW interface (<http://dot.imgen.bcm.tmc.edu:9331/multi-align/Options/map.html>), with subsequent editing in MS Word 97 for Windows.

Exon and gene structure predictions in repeat-masked were made using GENSCAN (Burge and Karlin, 1997) at the URL <http://bioweb.pasteur.fr/seqanal/interfaces/GENSCAN.htm>.

DNA and protein sequences were submitted to the MOTIF interface at <http://www.motif.genome.ad.jp> to search for putative transcription factor binding sites (TRANSFAC database) and amino acid motifs (Prosite Pattern and Prosite

Profile options). TFMATRIX entries are available by placing the accession number immediately after the "+" in the following internet address: http://www.genome.ad.jp/dbget-bin/www_bget?tfmatrix+.

The ORNL Grail form (<http://avalon.epm.ornl.gov/Grail-bin/EmptyGrailForm>) was used for prediction of CpG islands.

Neural network promoter predictions (NNPP) were performed at <http://www-hgc.lbl.gov/projects/promoter.html>.

Various protein characteristics were predicted by analysis through SAPS (Brendel *et al.*, 1992; http://www.isrec.isb-sib.ch/software/SAPS_form.html), HMMTOP (Tusndy and Simon, 1998; <http://www.enzim.hu/hmmtop/submit.html>) and TmPred (http://www.ch.embnnet.org/software/TMPRED_form.html).

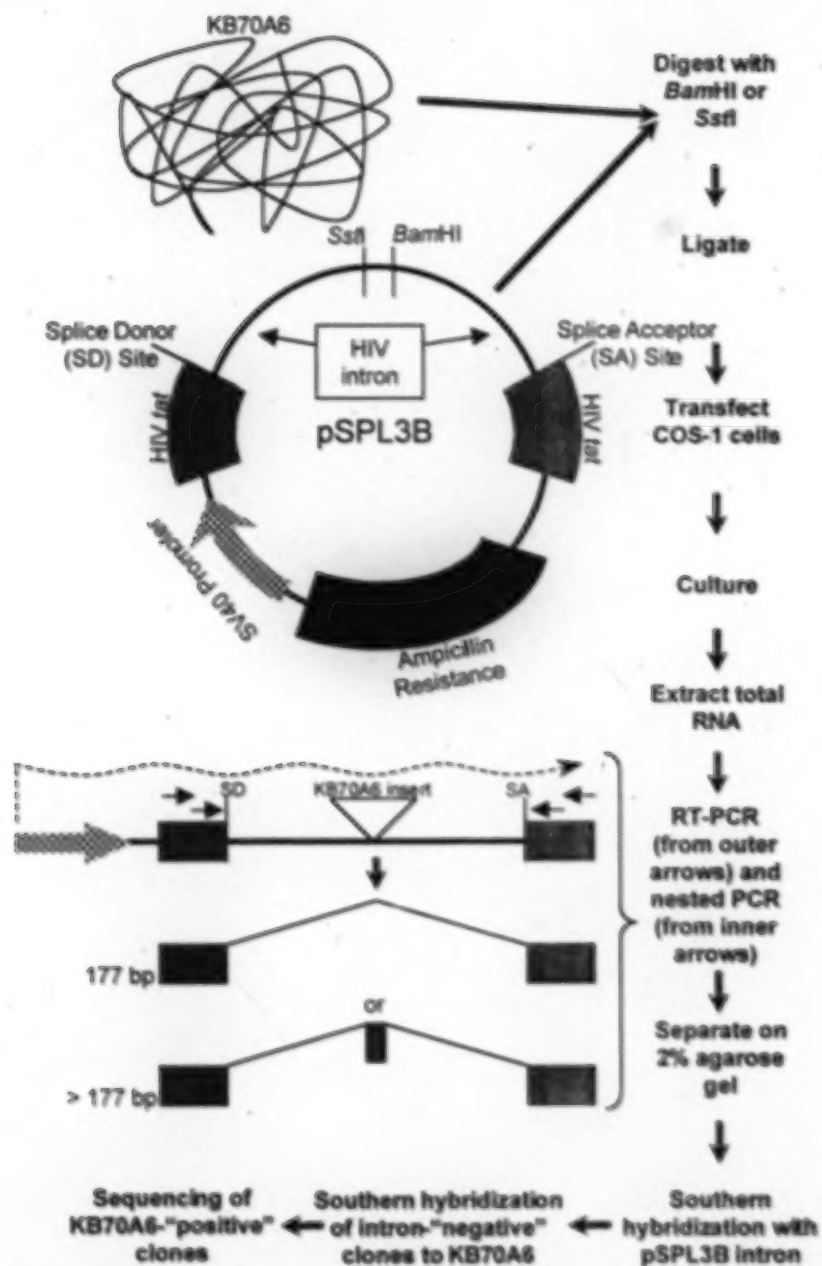


Figure II-1. Exon amplification of CESC BAC clone KB70A6.

The dotted line indicates *in vivo* transcription from the SV40 promoter (hatched arrow).

<u>Primer Name</u>	<u>Sequence (5' → 3')</u>
MBID1F	AGAAACGAGATGGACTGAGG
MBID1R	CTTCACGCTCATGGGCCGA
MBID2F	TAAGGAGAACAGCATTAGGAGA
MBID2R	GGAGGGAACATTTGCTTTTG
MBID3F	TCTTCCCTGGGTAACATCAC
MBID3R	TAGAGAAGAGCCTTTTATTTTG
MBID3R2	CTACCACTTAAGACAATTCACA
BIDENDF	AGCAAAGTGGTTCCTCTCTTA
BIDENDR	TTTATTTCCAAACAGTGGCCTC
MYOF	TCCTCCTTCTTGGGGAGCCA
MYOR	ATCTTGGCTTCAGCCATCCAC
M38F	CATTGGGGCTGGGGACCTAT
M38R	CCCTTTGTCAGCGGAGCCAT

Table II-1. DNA sequence of the PCR primers used in this study. All were obtained from Life Technologies, Inc. except for MYOF and MYOR which were engineered at the University of Alberta's Microbiology Department.

Chapter III: Results

Cloning, sequencing and expression of BID, a gene near the distal boundary of the CESC

To further delineate the CESC interval in humans, exon trapping (Buckler *et al.*, 1991) was performed on BAC clone KB70A6 (Figure III-1; Johnson *et al.*, 1999), to identify a gene that was beyond the distal boundary. This gene would then serve as an anchor on which to begin physical mapping of a CESC-homologous region in mouse. KB70A6 contains the marker D22S57, which defines the distal end of the interval duplicated in the dic r(22) CES patient (patient 25105; Mears *et al.*, 1995; Johnson *et al.*, 1999). Two exons ("70A6-240" and "70A6-380") were cloned and sequenced, then used to identify ESTs by BLASTN (Altschul *et al.*, 1990). When compared to the 5' EST for cDNA clone 52055, the longest clone identified at that time (insert size reported as 1128 bp), the exons appeared to be contiguous within the 5' end (Figure III-2A). Clone 52055 was obtained and fully sequenced to resolve ambiguities in the consensus derived from alignment of some of the most informative ESTs for this gene (Figure III-2A); the new consensus was submitted to GenBank under accession no. AF042083. Clone 52055 is recognized as the human mRNA for BID, a 195 amino-acid agonist of apoptosis. Conceptual translation of the cDNA sequence gave an exact match to the previously published amino acid sequence of human BID (Wang *et al.*, 1996). Using clone 52055 as a probe, Southern analysis of KB70A6, normal human genomic DNA, and genomic DNA from a human/hamster hybrid cell line containing chromosome 22 as the only human component demonstrated that *BID* is present as a single copy gene in humans, and confirmed its location on chromosome 22 (Figure III-3).

A BLASTN search with the *BID* cDNA identified numerous human and mouse entries in dbEST, of which the mouse clone 556749 was obtained and partially sequenced to confirm its identity. Northern analysis of adult mouse tissues with the *Bid* cDNA 556749 was consistent with the results of Wang *et al.*

(1996), showing an approximately 1.7 kb transcript in all tissues studied, but with greater abundance in kidney (Figure III-4B). However, the size of the insert for this cDNA was ~1.9 kb, which indicates that band size determinations on the Northern blot used can be inaccurate for up to 0.2 kb. The originating libraries of the human ESTs identified from the BLASTN search, including lymphocyte, placenta, tissues in the CNS and digestive systems, and several fetal tissues, indicate that *BID* is expressed in a variety of tissues. Hybridization of clone 52055 to Northern blots of human adult and fetal samples (Clontech, Inc.) revealed that *BID* is expressed in all tissues tested (Figure III-4A), but shows greater expression in the adult heart and liver. Unlike the expression in the mouse (Wang *et al.*, 1996), two messages were detected in each lane: an abundant transcript of around 1.2 kb and one at ~2.4 kb. The size of clone 52055 (1.1 kb) is consistent with it representing a near complete copy of the smaller human transcript. The 3' sequence of similar *BID* ESTs (Figure III-2A) indicate that some transcripts terminate ~150 bp earlier than 52055 (e.g. clones 627125, 128065 and 589582), which may explain the diffuse band in the Northern blot (Figure III-4A). The discovery of EST 503784-5' indicates that an mRNA with a further elongated 3' UTR may account for the ~2.4 kb message. The 5' end of clone 503784 overlaps the 3' end of clone 52055 by an identical 114 bp and would extend the transcript to more than 2.2 kb. As the genomic sequence was not yet available, it was unknown whether the 3' end of this cDNA was part of a true *BID* transcript or if the molecule was chimaeric. To confirm the localization of both ends of clone 503784 to the region, primers designed to its 3' end (BIDENDF and BIDENDR; Table II-1) amplified a 452 bp product from the original BAC KB70A6, but no other BACs in the region (data not shown).

Physical mapping of human BID outside of the CESC

The distal breakpoint of the dic r(22) (in patient 25105), which defines the current CESC, is located between *ATP6E* and *D22S57* (Mears *et al.*, 1995), but

the location of *BID* relative to this breakpoint was not known. Therefore Southern dosage analysis with exon 70A6-380 (211 bp) as a probe was performed. Results from pixel volume comparisons on a phosphorimaging scanner (Figure III-5; Table III-1) demonstrated that *BID* is present in three copies in a patient with trisomy 22 (GM07106), but only in two copies in patient 25105 and the normal diploid control. This evidence indicates that *BID* is not duplicated in patient 25105, although it must be within the region duplicated in all patients with a CES chromosome (Figure I-1).

To confirm that *BID* is overexpressed in a typical CES patient carrying a CES chromosome (Figure I-1), clone 52055 was used to probe a Northern blot of total RNA from lymphoblastoid cells of patient CM01, whose cells possess the supernumerary chromosome and therefore four copies of 22q11.2 (McTaggart *et al.*, 1999; Figure III-6; Table III-2). Dosage analysis using a phosphorimaging scanner showed that the patient cell line exhibits an approximately 1.8-fold increase in expression of the *BID* 1.2 kb transcript (compared to a normal diploid control line), while the 2.4 kb transcript was too faint to analyze.

Genetic mapping of *Bid* on mouse chromosome 6

The (C57BL/6JEi x SPRET/Ei)F1 X SPRET/Ei backcross panel ("Jax BSS panel") from The Jackson Laboratory (Rowe *et al.*, 1994) was used to localize *Bid* in the mouse genome, to investigate the conservation of linkage of CESC genes and their mouse homologues. PCR primers were designed from the 3' UTR sequence of overlapping *Bid* ESTs (GenBank accession nos. AA289104 and AA107402) in order to clone and sequence this region from both *Mus musculus* (C57BL/6JEi) and *M. spretus* (SPRET/Ei), the two species used in the backcross. Although closely related, the sequence of non-coding homologous regions (such as 3' UTRs) in these two genomes usually contain numerous differences. PCR primers were designed to amplify the complete ~1.2 kb 3' UTRs in three fragments (with primer pairs MBID1F/MBID1R, MBID2F/MBID2R

and MBID3F/MBID3R; Table II-1; Figure III-7). Each product was incompletely sequenced and 22 single-nucleotide variations were discovered. The combination of oligonucleotides MBID2F and MBID3R2 (Table II-1) amplified a 779 bp product from *M. musculus* which should contain only one *Dde* I restriction site (512 + 267 bp fragments), while a similar sized product from *M. spretus* should contain two sites (~423 + 267 + 89 bp fragments). PCR using genomic DNA from the 94 N2 progeny in the Jax BSS panel was performed and the products directly digested with *Dde* I, confirming the predicted restriction patterns (Figure III-7). The results were sent to the Jackson Laboratory for analysis. The haplotype comparisons indicated that *Bid* maps to a region on mouse chromosome 6, 55-60 cM from the centromere, that showed no crossovers with four other loci: *Apobec1*, *Rad52*, *Rny1*, and *D6Ggc2e* (Figure III-8).

The mouse homologues of three genes near the distal CESC boundary are linked on chromosome 6

In order to demonstrate conserved linkage of genes homologous to those in the CESC, putative mouse homologues were physically mapped to chromosome 6. A murine BAC library was hybridized simultaneously with three mouse cDNAs, representing *Bid* (clone 556749) and two novel genes whose human homologues map near *BID* (*CTCO* and *MTP*; G. Banting, unpublished). *CTCO* (3' end #1) was represented by clone 1003369 and *MTP* by clone 404173. The probes hybridized to BAC clones 67D14, 137D13, 141K23, 261P13 and 453L13. DNA was isolated, sized by pulsed field gel electrophoresis (PFGE) and tested by Southern blotting to confirm each probe's specificity. The clones overlapped, initiating the map in Figure III-9, thereby demonstrating close linkage of these three genes on mouse chromosome 6.

Prediction and mapping of 2 novel genes and IL-17R from human genomic sequence (PAC 143I13)

For the purpose of isolating mouse-specific probes that are homologous to genes on chromosome 22q11.2 proximal to *BID* (i.e. within the CESC), the quickest approach was to analyze human genomic DNA sequence made available through human gene cloning strategies for the CESC. Identification of CES candidate genes began with exon trapping in the McDermid lab (Riazi, 1998; Brinkman-Mills, 1999) and direct sequencing (by Dr. B. Roe at the University of Oklahoma) of a BAC/PAC contig across the CESC (Figure III-1; <http://www.genome.ou.edu/maps/ch22.html>). Annotation of the complete sequence of PAC 143I13 allowed the precise mapping of three additional genes on 22q11.2: *IL-17R*, *BTPUTR* and *PSL*. Submitting 143I13 to a BLASTN search of the nr database indicated similarity of the centromeric end to the 3' end of the gene for the *interleukin-17 receptor (IL-17R)*, making it the most proximal gene identified in the CESC. Interestingly, the mouse homologue (*Il-17r*) has been genetically mapped to chromosome 6 (Yao *et al.*, 1995), very close to the location of *Bid* (Figure III-10), suggesting that the human and mouse genomes share conserved synteny for the interval between these two genes.

Clusters of ESTs near the centromeric end of 143I13 (discovered through a BLASTN query of dbEST) suggested that there are alternate 3' UTRs for *IL-17R* (Figure III-11) incorporated by either alternative splicing or differential polyadenylation of pre-mRNAs. Two of the largest available cDNA clones (representing the different 3' ends; see below) were obtained and sequenced fully. There is no overlap between either of them, nor to the published incomplete mRNA sequence of *IL-17R* (Yao *et al.*, 1997). However, since the first canonical polyadenylation signal (AATAAA) distal to the stop codon of *IL-17R* in the PAC sequence is only 915 bp upstream of clone 310354, this cDNA (which does not contain a significant ORF) probably comprises part of an *IL-17R* transcript that uses a more distal signal, rather than representing an adjacent gene. The genomic interval between the 3' ends of clone 310354 and the EST

cluster represented by clone 366663 is only 2981 bp (Figure III-11). Due to the lack of any appreciable ORFs in this interval, clone 366663 probably represents an alternative 3' end of *IL-17R* rather than a separate gene. Preliminary results with RT-PCR have also suggested that these ESTs are part of *IL-17R* transcripts (S. Maier, unpublished). A query of dbEST with the mRNA sequence for mouse *Il-17r* resulted in identification of clone 303764 as a partial cDNA; sequencing of this clone verified its identity.

Human cDNA clone 46414 is one of the longest members of an EST cluster that suggests the presence of the novel gene *BTPUTR*, transcribed from telomere-to-centromere (i.e. on the opposite strand from *IL-17R*) whose 3' end is 550 bp downstream of clone 366663 (Figure III-11). Clone 46414 was obtained and fully sequenced, and demonstrates that the cDNA does not splice out any introns nor contains a significant ORF; it is therefore likely to be entirely within the 3' UTR of *BTPUTR*.

Distal to, but in the same orientation as, *BTPUTR* is another putative gene on 143I13, *PSL*, discovered by analysis of overlapping ESTs (Figure III-11). The sequence of cDNA clone 52444 was determined and allowed the characterization of several exon-intron boundaries at the 3' end of the gene.

Further elucidation of the structure of *BTPUTR* (see below) and *PSL* was aided by analysis with an additional cDNA for *PSL* (S. Maier, unpublished), homologous partial mouse genomic sequence (obtained later), gene prediction by GENSCAN, as well as for searches for homology to primary and secondary amino acid structures. In short, *PSL* is somewhat similar to phosphatidyl synthase from *Schizosaccharomyces pombe* while *BTPUTR* may encode a leucine zipper domain, found in some transcription factors. To obtain a probe for mouse *Psi*, PCR primers were designed to amplify 165 bp from an exon for which the corresponding cDNA (EST 1050354-5') was unavailable, but showed 84% identity over 172 bp with human clone 52444. The PCR product was sequenced to confirm its identity, prior to screening the mouse BAC library. EST AI429830 showed homology to *BTPUTR*, with 86% identity over 114 bp at the

very 3' end. It was used to search for a longer clone and identified cDNA 464798 which was obtained to serve as a probe for the murine homologue.

Physical mapping of 14 mouse genes (in a region homologous to human 22q11.2) onto an assembled BAC contig

In order to extend the mouse chromosome 6 BAC contig initiated by physical mapping of *Bid*, a series of further BAC library screens were performed, using mouse cDNAs discovered through sequence analyses and probes identified by other researchers characterizing genes in the CESC in the McDermid lab. The second library screening reaction, using a probe for *CTCO* (5' RACE #1; G. Banting, unpublished) that is more proximal than the partial cDNA used previously (see above) resulted in the identification of clones 52007 and 555D9 (Figure III-9). The next "batch" hybridization included cDNA probes for *Il-17r* and *Btputr* as well as the 165 bp *Pst*I PCR product. This reaction identified the remaining clones, except for 369P18 (Figure III-9).

Once the BACs were purified, digested (with *Hin* dIII, *Sst* I, *Xho* I or *Rsa* I) and blotted to membranes, successive hybridizations were carried out to determine the specificity for each probe used during library screening. The inserts were also sized by PFGE. This allowed for the organization of these clones into two non-overlapping sets of contiguous inserts. BACs 596K8, 541L22, 555D9, 67D14, 141K23 and 453L13 were selected for sequencing and delivered to B. A. Roe (Figure III-9; http://www.genome.ou.edu/mus_totals.html).

Once partial mouse genomic sequence was available, it was checked against the available corresponding human sequence with "BLAST 2 Sequences". With this method, a ~2.1 kb region of 541L22 showed 65.6% identity with human PAC 238M15 (Figure III-1). Such a high level of similarity for a fragment that shows no significant ORF suggests it is located in an UTR of a gene (*CES38*) for which a human exon was previously trapped (Brinkman-Mills, 1999). Upon amplification of 1.6 kb of this fragment from mouse BAC templates

with primers M38F and M38R (Table II-1), it was tested against the mouse BAC library and hybridized to clone 369P18 (Figure III-9) which was also submitted for sequencing.

By Southern analysis, the relative order along the contig for the seven genes discussed above was determined (Figure III-9). As well, eleven additional markers (including seven additional genes or gene fragments) were placed on the physical map in this fashion or by sequence analysis. Partial sequence for mouse BAC 596K8 demonstrated the presence of *Il-17r* as well as 85-95% identity to BAC clones (283I3 and 350L7) localized to human chromosome 12p13.3. A comparable level of identity was also exhibited for the gene *retinoblastoma binding protein 2 (RBBP2)* that is contained within these BACs (http://www.hgsc.bcm.tmc.edu/seq_data_old/cgi-bin/bcm-web-regions.cgi?region=12p13.3). Due to the large size of this gene (>50 kb in humans; determined by comparing the mRNA to BAC 283I3), it is unlikely to map to the "right" of *Il-17r* (Figure III-9). These results are not surprising as other homologues of genes from human 12p13.3 are genetically-linked to *Il-17r* (see Discussion and Figure IV-2). The gene *Idgfl*, for which human and pig ESTs are available, was localized by hybridization. The human cDNA clone 54445 was divergent enough that it did not hybridize to the mouse BACs, but the porcine clone F14844 (kindly provided by A. K. Winteroe) gave a weak but specific signal for the BACs shown in Figure III-9. A human exon ("11A") amplified from clone 238M15 (Brinkman-Mills, 1999) hybridized weakly to BACs 369P18 and 555D9; it is presumed to be part of a novel gene structure (*CES11*) distinct from the adjacent *CES38*. The next marker is a 5' RACE product (#2) for *CTCO* that is the most proximal probe available for the gene (G. Banting, unpublished). An alternative 3' end for mouse *Ctco* is represented by cDNA 1005753, which is conserved with the homologous region in humans (data not shown); this probe is labeled as "*Ctco* 3' end #2" in Figure III-9. *Atp6e* was mapped by hybridization with the human cDNA 61EW (Baud *et al.*, 1994) and with a human genomic fragment containing the 3' end of the gene (DD8; Baud *et al.*, 1994). *Gab* was predicted through comparative sequence analysis (see below). Its 5' end was localized with a 600 bp probe

from cDNA 1003631. The alternative 3' ends, "#1 and #2" in Figure III-9, were mapped only by sequence analysis, corresponding respectively to the 3' ends of cDNAs 1003631 and 1181972. Two additional genes localized to the CESC homologous region on mouse chromosome 6 are *Kiaa0819*, presumed to be the homologue of a fully sequenced human mRNA, and *Ng453*, a novel gene represented by mouse ESTs 1382292-3' and 516578-3'. Both of these genes were predicted through BLAST analysis of BAC 453L13, and their location relative to *Bid* places them beyond the region homologous to the distal boundary of the CESC. Human *KIAA0819* was mapped to a position distal to *BID* by Southern hybridization of cDNA clone 305358 to BAC 154H4 (data not shown; Figure III-1). This cDNA was identified by querying dbEST with the human mRNA sequence. The location of human *NG453* is unknown.

Prediction of a new gene (GAB/Gab) from human and mouse genomic sequence

Coordinated sequencing of human and mouse genomic clones near the distal CESC boundary between *ATP6E* and *BID* allowed for a comparative analysis approach to be undertaken for the characterization of *GAB*, a gene which might be duplicated in every CES patient. A 106 kb interval between the genes *ATP6E* and *BID* on PAC 1087L10 (Figure III-1) demonstrates numerous EST "hits" upon performing a BLASTN query. The vast majority of the ESTs agree with a hypothesis that this interval contains a single multi-exon gene with a complex transcription pattern including alternative splicing and alternative polyadenylation (Figure III-12).

Several of the *GAB* cDNAs were obtained and sequenced fully or partially. Clone 1541822 is a spliced product that may represent the shortest transcript of *GAB*, containing exons 1-7 and utilizing a stop codon and polyadenylation signal within 22 bp immediately downstream of exon 7. This transcript terminus is supported by numerous other ESTs suggesting that it is not an artifact of library

construction. All of the other transcripts presumably splice exon 7 to the large exon 8 thereby increasing the size of the predicted protein product by 163 amino acids. Northern analysis using clone 1541822 identified at least five distinct transcripts for *GAB* (approximately 1.3, 2.4, 4, 5 and 10 kb) in all tissues tested (Figure III-13) plus brain, thymus, spleen, kidney, liver, placenta, lung, leukocytes, testes and ovary (data not shown). The 2.4 and 5 kb bands were consistently weaker than the others. Transcripts of ~1.3, 4, 5 and 10 kb can be constructed by differential usage of exon 8 and of those polyadenylation signals in the 35 kb downstream of the stop codon in exon 8 that each correlate to the 3' ends of numerous ESTs. This theory assumes that clone 1541822 is a nearly full-length cDNA for the 1.3 kb mRNA (Figure III-12). The structure of the 2.4 kb message remains elusive. Fifteen putative polyadenylation signals (AATAAA or ATTAAA) were identified downstream of the exon 8 stop codon but before the stop codon for *BID*. Since 15 transcripts longer than 1.3 kb were not detected by Northern analysis it is possible that: 1) some of the elements are not truly functional, 2) certain Northern blot bands represent two or more similarly-sized transcripts, or 3) certain signals are used infrequently and thus could not be detected by hybridization, or some combination of these scenarios.

The abundance of EST hits for *GAB* suggests it is a highly-expressed gene. As such, it was possible to select several partial cDNA clones each representing the 3' end of different messages. Of these, clone 1854295 was obtained, sequenced fully and used for Northern analysis (Figure III-13). It is clear by its sequence and hybridization pattern that this probe encompasses the terminal 1011 bp of the largest *GAB* transcript of ~10 kb. Careful analysis of EST directionality at the junction of *GAB* and *BID* reveals that the longest transcripts of each gene overlap over a 1077 bp interval entirely within 3' UTRs (Figure III-2B), and therefore no part of cDNA 1854295 is unique to *GAB*. Canonical polyadenylation signals are located ~1.1 kb apart in opposite orientation on the two DNA strands, in positions corresponding to the termini of the overlapping ESTs (see Figure III-2B). On the Northern blot, 1854295's detection of a 2.4 kb band likely represents hybridization of the double-stranded DNA probe to the

longest transcript of *BID* (Figures III-12 and III-13), although it may recognize the *GAB* 2.4 kb molecule as well. This band is much more prominent than when the blot was probed with the first *GAB* cDNA.

Based on EST clustering, murine expression of *Gab* appears superficially less complex than for *GAB*. Firstly, there are only two different transcripts for *Gab*, one (2.4 kb) terminating just beyond the region homologous to human exon 8, the other (6.2 kb) terminating much further downstream, adding ~3.8 kb solely to the 3' UTR. cDNA clone 1003631 (see physical mapping results above) was obtained and the 2.4 kb insert sequenced to reveal the organization of exons 1-6, and the short form of the terminal exon 7 (Figure III-12). This clone's terminus is just upstream of a polyadenylation signal that is likely used for the natural 2.4 kb transcript (i.e. 1003631 does not contain a poly-A tail so is probably a truncated partial cDNA). The transcript with the extended 3' UTR is represented by clone 1181972 (see previous physical mapping results); it was obtained and its 3.7 kb insert was sequenced at both ends to confirm its identity. Partial restriction mapping also suggested it corresponds to 3.7 kb of intronless genomic sequence immediately following the first 3' end (data not shown). The longer *Gab* message does not overlap *Bid* (which produces only one mRNA). Two conserved, possibly functional, elements appear in the longer 3' UTR of human *GAB* (Figure III-12). One element (119 bp), labeled "p" in Figure III-12, contains the polyadenylation signal of the 2.4 kb mouse transcript and matches the end of the 4 kb human transcript, while the other ("q"; 50 bp) is near the end of the 6.2 kb mouse mRNA and matches the end of a possible 7 kb human mRNA. The coding DNA sequence of *GAB* (human exons 3-8) displays 74.1% identity to *Gab* (mouse exons 2-7), but there is no significant homology in the UTRs apart from the domains discussed above. The characterization of *Gab*'s 5' UTR is ongoing, and should be made clearer once sequencing of mouse BAC 141K23 is complete. Examination thus far has not revealed homology between human *GAB*'s first and second exon with the corresponding region of the mouse genome (on BAC 141K23). There is however a nontranscribed segment just distal to *GAB* exon 2 that corresponds to exon 1 in mouse (homologous region "a"). The

reason for this sequence to be transcribed in mouse but not in humans is unknown.

In Figure III-12, the homologous fragments labeled "a" to "t" (in pink) were identified by a "BLAST 2 Sequences" comparison of 133 kb from repeat-masked PAC 1087L10 and 81 kb from mouse BACs 141K23 and 453L13. This human region contains sequence from just upstream of *GAB* to the end of PAC 1087L10, centromeric to *BID* exon 1. The mouse fragment contains sequence from a similar distance upstream of *Gab* exon 1, to *Bid* exon 1. The sequence of 141K23 is incomplete, but separate contigs were assembled by comparison to the human sequence. The homologous regions ranged from 140 to 355 bp in length for those that contained protein-coding sequence, and from 36 to 197 bp for "noncoding homologous segments" (NCHs). The relative spacing of all of these elements appear conserved, although *Gab* is more compact in mouse (106 kb in humans; 57 kb in mouse, not including a span corresponding to the large interval from human exons 1-2). In total, ten elements contain *GAB* and *BID* protein-coding exons (and some flanking intron sequence), while eleven are NCHs located in introns and UTRs. The seven coding homologous segments for *GAB/Gab* ("d, f, g, i, m, n and o") have a combined length of 1376 bp and 1362 bp in human and mouse, respectively, at an overall level of 83.3% identity. The ten NCHs for *GAB/Gab* ("a, b, c, e, h, j, k, l, p and q") total 1136 bp and 1137 bp in length (human and mouse, respectively), at an overall level of 81.7% identity. The NCHs comprise 1.1% of the genomic region of *GAB* and 2.0% of the *Gab* region. The NCHs were submitted separately to MOTIF for comparison against the vertebrate section of the TRANSFAC database of transcription factor binding sites. Numerous matches and near-matches to the consensus recognition sequences were identified. Among those elements that perfectly fit the canonical sites, only two were conserved at homologous positions in human and mouse. The sequence TGAGGGGA appears in both the human and mouse homologous regions "a". This matches the consensus NGNGGGGA for the binding of the myeloid zinc finger protein (MZF1; TFMATRIX accession no. M00083). In homologous region "e", the human sequence TGCATTTAATTAATCC fits the

consensus of (A/T)NNAN(C/T)(C/T)AATTAN(C/T)NN for binding of the S8 homeodomain (TFMATRIX accession no. M00099). The corresponding mouse sequence is TTCATCTAATTAATCC. The matches to highly degenerate consensus sites were not studied as many were probably identified by incidental similarity. The 5' regions of *GAB* and *Gab* were also examined for predicted CpG islands (by Grail) and promoters (by NNPP). A high scoring promoter (0.98 out of 1.0) was predicted just upstream of *GAB* exon 1, but in the reverse orientation. This also places it just upstream of *ATP6E* which is transcribed from the opposite DNA strand. No promoters were predicted for *Gab*. Homologously-situated CpG islands were predicted around region "a" (Figure III-12). A CpG island was also predicted surrounding the *ATP6E* promoter.

To characterize this novel gene further, the conceptual translation products of human cDNAs were determined. They predict two different forms of *GAB* protein. One form is 201 amino acids long, translated from mRNAs that do not contain exon 8. The longer ORF lacks a single residue coded by exon 7 (due to alternative splicing), and is replaced by 285 amino acids, translated from those molecules containing exon 8. The longer sequence (485 residues) and the predicted ORF for mouse *GAB* (434 residues) are 69.1% identical, and are aligned in Figure III-14. The two proteins demonstrate high conservation over most of the sequence, but only weak similarity in the carboxy-terminal portion. BLASTP queries to the nr database did not reveal homology to any known proteins. The three amino acid sequences were also analyzed by online programs for possible functional domains. The short human isoform was unremarkable, but the other human protein and the mouse protein were predicted (by HMMTOP) to contain a transmembrane domain at the carboxy-terminus, with the bulk of the protein located inside the cell. The membrane-spanning region was also predicted by TmPred. Internal repeated stretches of amino acids were discovered by Statistical Analysis of Protein Sequences (SAPS) in each of the long human and mouse ORFs. The element EELKSLD is encoded by human exon 4 and mouse exon 3. EEVKSLD is encoded by human exon 8 and mouse exon 7, allowing the two domain pairs to fit a consensus of

EE(LV)KSLD (Figure III-14). Also encoded by exons human-8 and mouse-7 is the element SWQSESLPVSL, followed by a four-residue gap, and then the element SWHTESLPVSL (consensus = SW[QS/HT]ESLPVSL). No functional activity is attributed to these repeated domains, however all four have been perfectly conserved in both species.

Annotation of genes and gene duplications in the CESC and mouse chromosome 6

As a step towards discovering and characterizing new candidate genes in the CESC, and examining the utility of homologous genome sequences to aid this process, other regions of the CESC were analyzed for gene content and evolutionary conservation. *ATP6E* is the next proximal gene to *GAB*, in both humans and mouse, and is 93% identical at the protein level to *Atp6e*. 36 kb of incomplete human genomic sequence containing all of the exons of *ATP6E* was compared to 29 kb of nearly-complete mouse sequence, containing all of the exons of *Atp6e*. Thirteen homologous elements were identified in total (Figure III-15). Nine of these contained protein-coding exons (and flanking intron sequence), totaling 782 and 780 bp (human and mouse, respectively) at an overall level of 91.7% identity. The four NCHs had a combined length of 149 bp (in each species) with 91.9% identity. Contrary to the organization of *GAB*, the NCHs within *ATP6E* contribute to ~0.5% of each genomic region and are confined to the extreme 5' and 3' ends (regions "a" and "m") and the small regions "c" and "d" in the first intron. Searching the TRANSFAC database with the NCHs resulted in identifying only one transcription factor binding site that occupied homologous positions. In region "a", the sequence CTTTATA (same in both species) fits the consensus recognition sequence of (C/A)TTTAT(A/G) for the chicken *CdxA* homeobox gene (TFMATRIX accession no. M00100). CpG islands were also predicted at the homologous 5' ends (by Grail), but NNPP predicted a promoter for only the human gene (see analysis of *GAB* above).

The next proximal gene to *ATP6E* is *MTP*, for which human cDNA and RT-PCR products have been sequenced (G. Banting, unpublished). It contains eleven exons, but the first three did not identify homologous sequences in mouse BAC 67D14 (Figure III-16). Murine cDNA clone 404173 was partially sequenced at both ends to reveal the organization of most of the mouse exons. To check the validity of a gene prediction program's characterization of a mouse gene for which incomplete cDNA information was available, 25 kb of DNA encompassing *Mtp* was submitted to GENSCAN. It was successful in predicting exons that correspond to exons 3-7 of the mouse cDNA, as well as elucidating the structure of exons 8-10 (Figure III-16). Only portions of exons 8 and 10 were sequenced from the cDNA, so it was unknown how long they were or if exon 9 existed. The GENSCAN-prediction of exon 9 gains further support from two further analyses. First, it corresponds to a region ("i") that is homologous to human coding exon 10, and second, its inclusion in the putative mouse ORF correlates with the homologous human protein sequence. GENSCAN also predicted two false-positive exons: one lies just upstream of exon 2 (determined from the cDNA) and the other lies adjacent to a sequence contig break (which may contain low-quality sequence on its flanks until the gap is covered). As a comparison, the human region was also submitted to GENSCAN, which correctly predicted exons 2-11, but made incorrect predictions at the 5' end (Figure III-16). There is a striking lack of homology in the 5' half of the genomic regions. Whereas translation is predicted to initiate in mouse exon 2, the human protein likely begins in exon 3 (G. Banting, unpublished). The residues coded by the remaining downstream exons in each species demonstrate rather high conservation of 88.0% identity over 317 amino acids. Analogous to the genomic context of *ATP6E/Atp6e*, the *MTP/Mtp* regions contain few NCHs. Only two NCHs were discovered (regions "a" and "b"), which have a combined length of 165 bp in human and 163 bp in mouse, corresponding to ~0.6% of each genomic region. Interestingly, the "a" fragments both fall upstream of the first exons, while the "b" fragments are both located in the intron following the translation initiation codon. These NCHs share 79.4% identity overall, while the protein-coding homologous regions share 74.6%

identity overall (1264 and 1096 bp combined lengths in human and mouse). One conserved putative transcription factor binding site was found in the NCHs, in region "b". The sequence AGATAAGAACA (same in human and mouse) matches the consensus recognition site for the family of GATA-X binding proteins, NGATAAGN(A/C)NN (TFMATRIX accession no. M00203). No CpG islands or promoters were predicted for these genes. The protein products appear to encode brain and liver-specific mitochondrial membrane transport proteins (G. Banting, unpublished). As GATA-binding proteins are implicated in regulation of cell-restricted gene expression profiles (Suzuki *et al.*, 1996), it is enticing to propose that the NCHs of *MTP/Mtp* specify its transcription in brain and liver only, perhaps mediated by a novel GATA factor(s).

Characterization of the gene *BTPUTR* is a complex problem for which homologous mouse sequence has solidified a prediction of the protein sequence. Although there are numerous EST hits for the ~20 kb interval of human PAC 143113 between *IL-17R* and *PSL* (Figure III-11) none appear to be derived from spliced products nor do they code for ORFs of significant length (>100 amino acids). The sequence of cDNA clone 46414 is illustrated in Figure III-17, relative to predicted features of the region. It is ~2 kb in length, and its continuity with the genomic sequence suggests it represents part of or all of the 3' UTR. Predicted proteins of >50 amino acids (i.e. from methionine to STOP) are shown in the three reading frames in the telomere to centromere orientation in Figure III-17 (corresponding to the direction of transcription of cDNA 46414). The first polypeptide in frame -3, from the right, is the longest uninterrupted coding sequence in the area (578 amino acids). All three frames contain long stretches of ORF that could be spliced together to create a larger product, however, GENSCAN predicted *BTPUTR* to be a single-exon gene, with an ORF corresponding to the 578-residue product mentioned. This result was weighed with the fact that frame -1 also contains a significant continuous amino acid sequence of 209 amino acids. Neither of the conceptual translation products detected significant homology to proteins in the nr database (by BLASTP). The DNA sequence surrounding the proposed translation initiation codons was

analyzed for similarity to the Kozak consensus CC(A/G)CCATGG, in the 5'-3' orientation (the methionine codon is underlined; Kozak, 1987). The region encoding the 578-residue ORF (frame -3) contains the sequence GGACAATGC (5 of 9 nucleotides match) while the region encoding the 209-residue ORF (frame -1) contains GCCGGATGC (4 of 9 match). Therefore, neither sequence convincingly indicates the actual translation initiation site. The GC-richness of the region is reflected by the prediction of a large CpG island overlapping the candidate ORFs. The NNPP-prediction of a promoter in this island suggested transcription begins just upstream of the frame -3 ORF. The most compelling evidence that the GENSCAN-predicted ORF in frame -3 encodes the true BTPUTR product, comes from homology to partial mouse genomic sequence. A portion of mouse BAC 541L22 that overlaps the region that would encode the amino-terminus of the 578-residue BTPUTR product was examined for homologous ORFs. Unfortunately, a contig break within the mouse sequence allowed for only ~500 bp to be analyzed, and did not include the region homologous to the frame -1 ORF. Nonetheless, only one frame in the mouse contains a methionine residue (at a site homologous to the human frame -3 initiator methionine), and the protein product has 85.4% identity with the 578-residue human ORF, over a stretch of 36 amino acids. Partial ORFs in the other two mouse reading frames have <55% identity with corresponding human ORFs. A search of the PROSITE database with the 578-residue human sequence revealed the presence of a putative leucine zipper motif (L-PAHLRY-L-LIAVYF-L-TLASPV-L; PROSITE accession no. PS00029), which suggests BTPUTR may be a transcription factor that acts as a dimer (O'Shea *et al.*, 1989). On the other hand, TMPred's "strongest" model predicted nine transmembrane domains for BTPUTR, with HMMTOP confirming six of the membrane-spanning regions; both programs disputed the orientation of some of the domains. This contradicts the idea that BTPUTR is a nucleoplasmic transcription factor. NCHs identified with the partial homologous mouse sequence have a combined length of 761 bp (human) and 737 bp (mouse), at an overall level of 81.7% identity. A single partially-coding element of 140 bp was identified (covering the translation

initiation site) and shows 88% identity; this homologous segment is interrupted by a contig gap in the sequence of mouse BAC 541L22. Querying TRANSFAC with the NCHs did not identify conserved transcription factor binding sites. The legitimacy of the single-exon, 578-amino acid theory for *BTPUTR* should be clearer when sequencing of the homologous mouse region is completed.

Over 400 kb of DNA centromeric to *IL-17R* was assembled into four contigs of human genomic sequence (clones 15J16, 87O8, 20K14 and 109L3; Figure III-1). It was analyzed for additional CES-candidate genes which may have homologues mapping near *IL-17r* on mouse chromosome 6. No homology was discovered between this region and the incompletely-sequenced mouse BAC 596K8 that contains *IL-17r*. As illustrated in Figure III-18A, the *IL-17R*-proximal region is constructed of duplications of other regions of the human genome (see also Appendix II). Conversely, some of the regions could be duplicated from 22q11. The chromosome 22 sequence was repeat-masked before submission to BLASTN for comparisons with the nr and htgs (unfinished genomic sequence) databases. Therefore, the graphical representation of the paralogy shows numerous gaps (in Appendix II), where alignment was impossible due to the presence of masked interspersed repeats. RepeatMasker analysis revealed that 50.7% of the proximal 400 kb consists of SINEs, LINEs, LTR elements and repeated "DNA elements". LINEs alone account for 31.4% of the region, reflecting the relatively low GC-content of the region (40.6%). Non-aligned segments may also have been due to sufficient divergence between chromosome 22 and the duplicated regions (analysis not complete). The percent-identities indicated in Figure III-18B and Appendix II are those calculated by BLAST for the highest-scoring segment of each alignment, and is not representative of the entire alignment, although there was little deviation from this number for the majority of the scoring segments (data not shown).

This 400 kb region exhibits numerous paralogous gene segments, including exon-intron organizations conserved from the ancestral loci. Of note are the *von Willebrand Factor* pseudogene (at position 5-25 kb; derived from 12p13.3) and pseudo-*IgK* loci (245-280 kb; derived from 2p11), previously known

to map to this area of 22q11. Several of the alignments reveal that some duplicated segments are also present on additional chromosomes. This confounds an interpretation of which is the ancestral locus, as the percent-identity between chromosome 22 to each of the other regions is comparable. For example, the majority of the region from 70-185 kb is homologous to sequences from both chromosomes 19 and Y, at levels of 83% identity to 19, and 83-87% identity to Y. These chromosomes, in turn, are 83-88% identical to each other. The actual extent of each duplication is unknown as the corresponding regions of chromosomes 19 and Y are incompletely sequenced. The current status of these sequencing projects inhibits analysis of the actual level of conservation between 22q11 and the other loci. When completed, detailed examinations of percent-identity and the location of paralogous interspersed repeats should disclose enough information to determine the time-scale between each duplication event. A similar situation appears in the region from 210-285 kb, which reveals the fact that *IgK* pseudogenes exist on both 22q11 and chromosome 10. It is evident from the surrounding sequence that the pseudogenes were first transposed to either 22 or 10 (intact with introns and intergenic sequence), and then a larger region was duplicated to the other chromosome. It is also evident that an *IgL* pseudogene was also transposed (from 22q12) to position ~210 kb early enough to be included in the larger duplicated region shared with chromosome 10.

Chromosome 22-derived cDNAs have been identified in the region from 365-410 kb. This 45 kb portion is shown in Figure III-18B, where the constructed mRNA (and green exons) represents the gene *SAHL*, discovered by another researcher in the McDermid lab (Riazi, 1998). Very little of the area contains genomic sequence that is specific to chromosome 22, with contributions from chromosomes 1, 12, 16, 19 and 21 organized in (mostly) nonoverlapping domains. Portions of the chromosome 21 paralogy even contain >100 bp stretches with 100% identity. However, a unique combination of exons derived from separate duplications results in transcripts that are specific to the proximal CESCO region. These transcripts, which are characterized by alternative

splicing to different 3' exons suggest that a new gene was created by "shuffling" of exons with different chromosomal origins. Representative ESTs from clusters discovered by BLASTN comparison of the repeat-masked 45 kb to dbest are shown in Figure III-18B (the chromosome 22-specific ESTs were not included). EST AA677577 likely represents a gene located on chromosome 16 in the region of the genomic clone AC003034, owing to the extremely high (99%) similarity between these two sequences. Curiously, this EST demonstrates that 22q11 contains sequences homologous to two exons utilized in this cDNA, but only one of them is incorporated into *SAHL*. Sequence divergence has significantly altered the region within and flanking the "unused" exon on chromosome 22, rendering it nonfunctional. Two possible ORFs exist for *SAHL*, each >200 amino acids long (data not shown). Further characterization of *SAHL* involved GENSCAN and CpG island predictions which are indicated on Figure III-21B. They suggest that the 5' end of the gene falls in the direct vicinity of the 5'-most exon discovered thus far. However, GENSCAN and cDNA analyses fail to unequivocally determine the ORF encoded by this gene. The lack of homology to the mouse BAC contig (suggested by the failure of hybridization of *SAHL* cDNA probes and by the disruption in conserved linkage near *11-17r*) impedes characterization by a comparative sequence analysis approach such as the one implemented for *BTPUTR*.

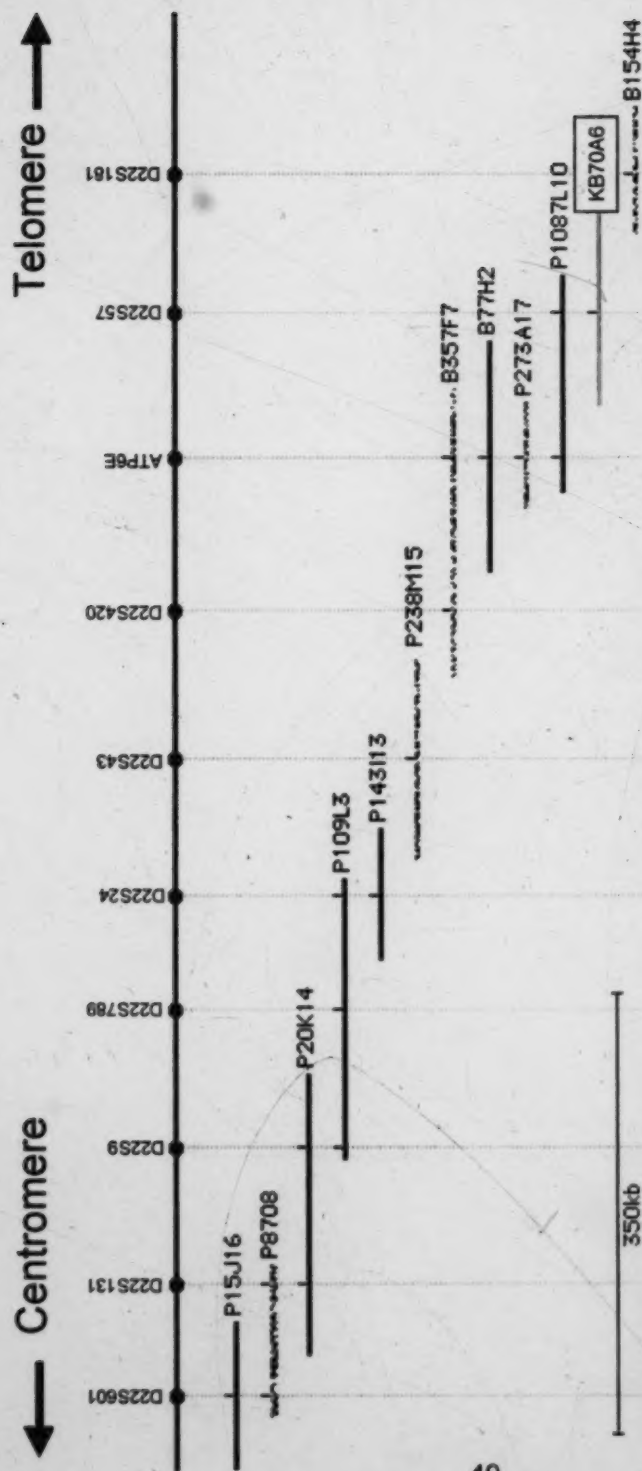


Figure III-1. BAC KB70A6 in relation to the genomic clones of the CESC clones being sequenced. Solid black bars indicate completed sequences, as of August 20, 1999, while stippled bars represent unfinished clones (KB70A6 is not being sequenced). The status of each clone's progress is listed at http://www.genome.ou.edu/hum_totals.html. STS and known gene markers are shown above the physical map. The clone sizes are drawn to scale, but the extent of overlapping is not necessarily accurate. This figure was adapted from Figure 1 of Johnson *et al.* (1999).

Figure III-2. Analysis of the *BID* transcription unit(s). **(A)** The sequence of clone 52055 submitted under GenBank accession no. AF042083 (the top, unshaded line in each block) was re-confirmed from the alignment of this clone's full sequence and the ESTs and exons listed. Identical residues are shaded in black, with mismatches shown in gray. The thymidine residue highlighted in green (just beyond the coding region) did not fit the consensus, but was unambiguous in the cDNA clone sequence; it may represent a polymorphism in the general population. The genomic sequence of PAC 1087L10 corresponding to all of the coding exons of *BID* is shown, with the introns removed (this clone does not contain the 5' end of the transcript). Arrows mark the positions of the introns. cDNA clone 128065 appears to be an alternatively-spliced product (a bracketed arrow marks the alternate position of the last intron) containing an extra 15 bp thereby adding five amino acids at the carboxy-terminus of the protein. Putative polyadenylation signals are shown in boxes. **(B)** Clustering of transcript termini for *BID* mRNA suggests three alternative 3' ends. The full mRNA (shown as the scaled gray bar) was assembled from the sequences of cDNA clone 52055 and the portion of PAC 1087L10 corresponding to the ~1.1 kb interval from EST 503784-5' to 503784-3' (dark blue bar). The location of the protein-coding ORF is shown as a hatched bar. Representations of the exons and ESTs from (A) are color-coded, and aligned to depict the 1.2 and 2.4 kb transcripts detected by Northern analysis (Figure III-4), and a putative ~1.0 kb message. The overlapping 10 kb transcript of *GAB* (detected by BLASTN) is discussed later.

(B)

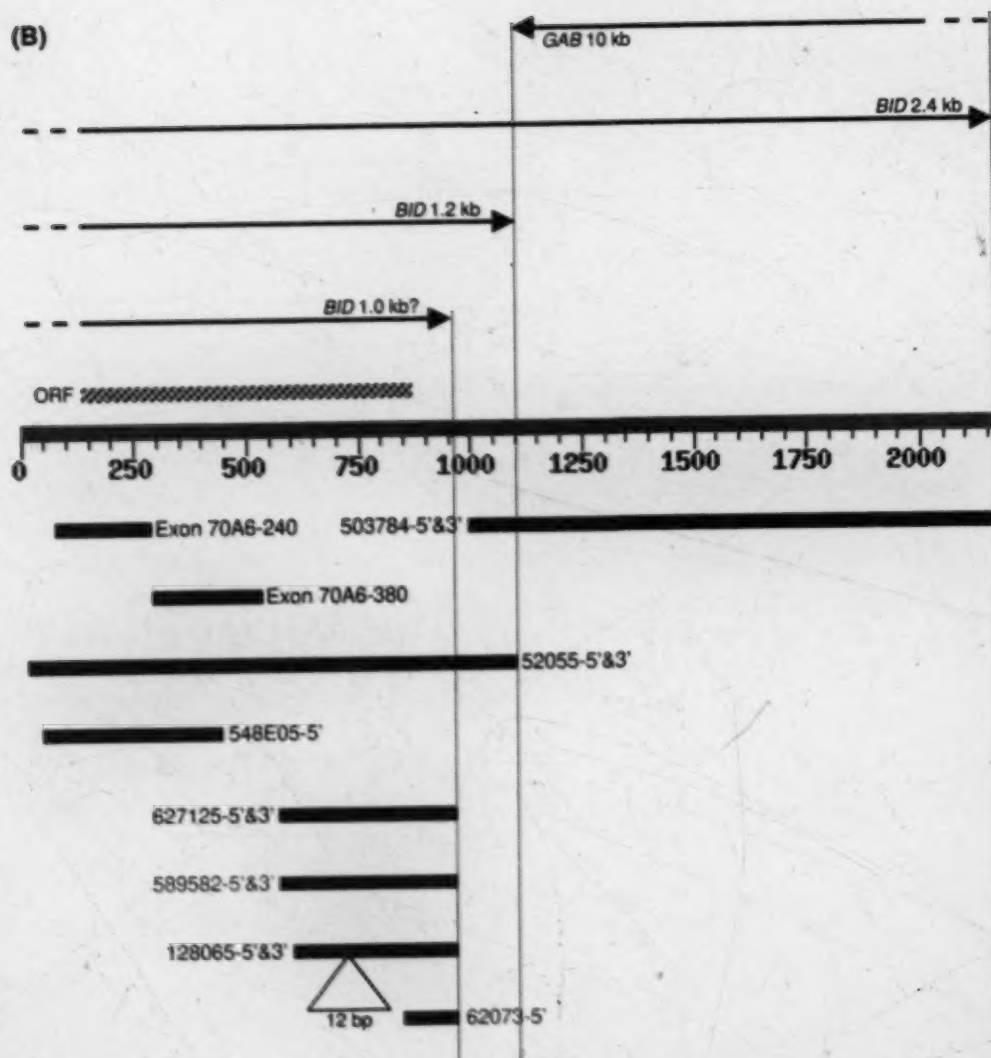
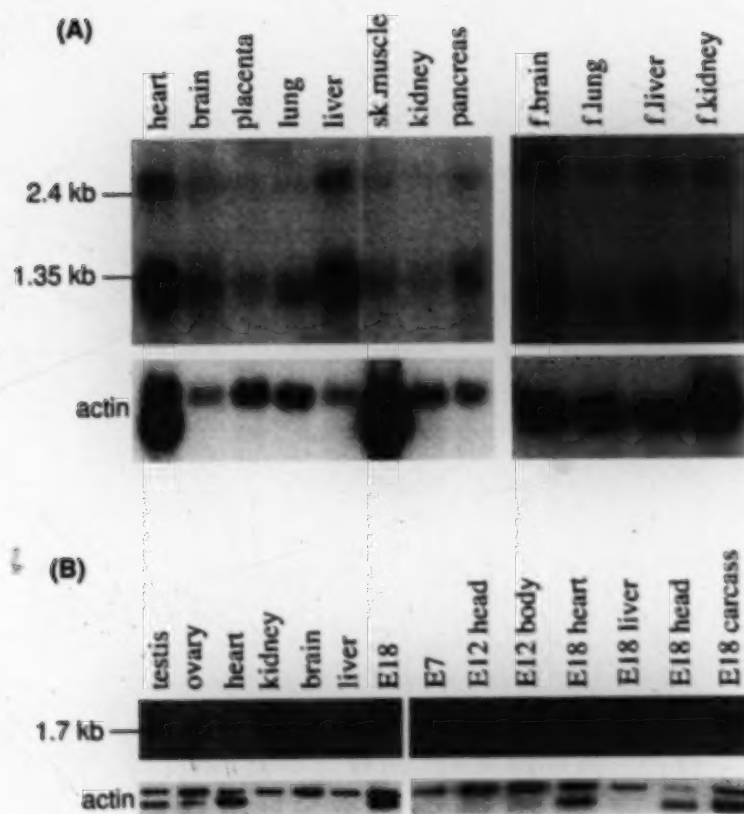




Figure III-3. Southern hybridization of *BID* to chromosome 22. The lower band in lanes 1 and 2 represent cross-hybridization of exon 70A6-380 to the hamster homologue of *BID*, while the upper band in lanes 1 and 3 is specific for sequences on human chromosome 22.

Figure III-4. Northern analysis of human *BID* and mouse *Bid*. **(A)** Commercial human adult and fetal multiple tissue blots (Clontech, Inc.) hybridized with cDNA 52055 (sk. = skeletal, f. = fetal). The adult blot was washed for 20 min at low stringency (25°C) and for 50 min at high stringency (50°C), before exposing for six days at -70°C. The fetal blot was washed for 20 min each at low (25°C) and high (50°C) stringency, before exposing for four days at -70°C). **(B)** Total RNA from adult and fetal mouse tissues (strain CD1; blots kindly provided by M. A. Riaz) probed with cDNA 556749. Both blots were washed for 20 min each at low (25°C) and high (50°C) stringency, before exposing for six days at -70°C. Fetal tissues are indicated by an E followed by the gestational age in days. E18 carcass = body without the head, heart and liver. In all panels, comparison to the signal from β -actin is shown for normalization. This probe cross-hybridizes to α -cardiac and α -skeletal actin (smaller band). The presence of this band in mouse adult testis and ovary is assumed to be due to contamination of the tissue with skeletal muscle.



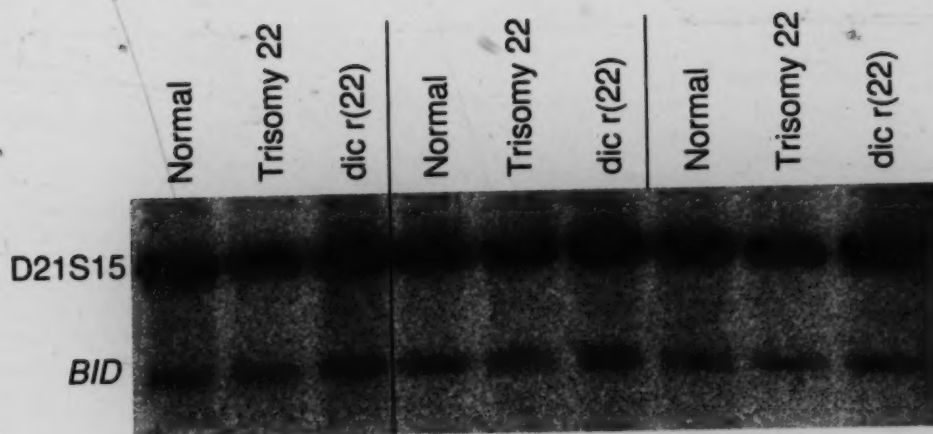


Figure III-5. Dosage analysis of *BID* for the CES patient with the smallest duplication. Radioactively-labeled exon 70A6-380 of *BID* was hybridized simultaneously with the reference probe D21S15 (from chromosome 21) to three replicates of normal DNA (two copies of 22q11.2), DNA from patient GM07106 with trisomy 22 (three copies) and DNA from the test subject (patient 25105) with the dic r(22). As shown in Table III-1, the trisomy 22 patient's dosage ratio is ~1.5-times greater than the diploid control while the dic r(22) patient's ratio was comparable to the diploid values.

	<u>Normal</u>	<u>Trisomy 22</u>	<u>dic r(22)</u>
1. D21S15 Signal Volume :	228951.59	143944.66	299778.18
BID Signal Volume :	42588.2	35954.87	48471.56
Dosage Ratio (BID+D21S15):	0.186	0.250	0.162
2. D21S15 Signal Volume :	216081.34	146777.42	242515.23
BID Signal Volume :	36084.66	33915.53	43880.6
Dosage Ratio :	0.167	0.231	0.181
3. D21S15 Signal Volume :	204859.57	129966.21	211918.92
BID Signal Volume :	36967.27	33190.92	41601.4
Dosage Ratio :	0.180	0.255	0.196
Ratio Average :	0.178	0.245	0.180
Intensity Relative to Diploid = (Ratio ÷ Normal) :	1.0	1.376	1.011
# of Copies of BID per Genome = (Relative Intensity x 2 Copies per Diploid Genome):	2	2.752 (3)	2.022 (2)

Table III-1. Calculation of *BID* dosage in the CES patient with the smallest duplication. Analysis of the hybridization intensities shown in Figure III-5 was performed by ImageQuant resulting in the "signal volumes" shown above. Each genome possesses only two copies of the nonsyntenic reference probe D21S15, located on chromosome 21. The trisomy ratios are significantly different (while the dic r(22) values are not different) from the control ratios, as determined by the Wilcoxon rank sum test (as described in Mears *et al.*, 1995).

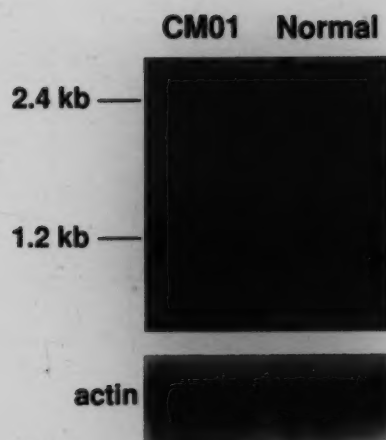


Figure III-6. Overexpression of *BID* in a typical CES patient with four copies of 22q11.2. Total RNA from the lymphoblastoid lines of patient CM01 (containing a CES chromosome) and a normal control individual were probed with radiolabeled cDNA 52055. Refer to Table III-2 for pixel volume analysis revealing an ~1.8-fold increase in expression of the 1.2 kb transcript in the patient cell line. An actin control probe was hybridized for normalization of the signals.

	<u>Normal</u>	<u>CM01</u>
1. <i>BID</i> 1.2 kb Signal Volume :	34171.72	56221.64
Actin Signal Volume :	4295337.91	3859363.32
<i>Dosage Ratio</i> (<i>BID</i> ÷ <i>Actin</i>) :	0.00796	0.01457
2. <i>BID</i> 1.2 kb Signal Volume :	30841.54	91438.75
Actin Signal Volume :	3089381.98	5004372.28
<i>Dosage Ratio</i> :	0.00998	0.01827
<i>Ratio Average</i> :	0.00897	0.01642
<i>Intensity Relative to Diploid =</i>		
(<i>Ratio</i> ÷ <i>Normal</i>) :	1.0	1.831

Table III-2. Calculation of relative *BID* 1.2 kb transcript levels in normal and CES patient cell lines. ImageQuant analysis of the signals depicted in Figure III-6 and one other replicate experiment (data not shown) demonstrate an average of 1.831-fold more expression of the 1.2 kb *BID* mRNA in a patient (CM01) with four copies of the gene, relative to a normal diploid control.

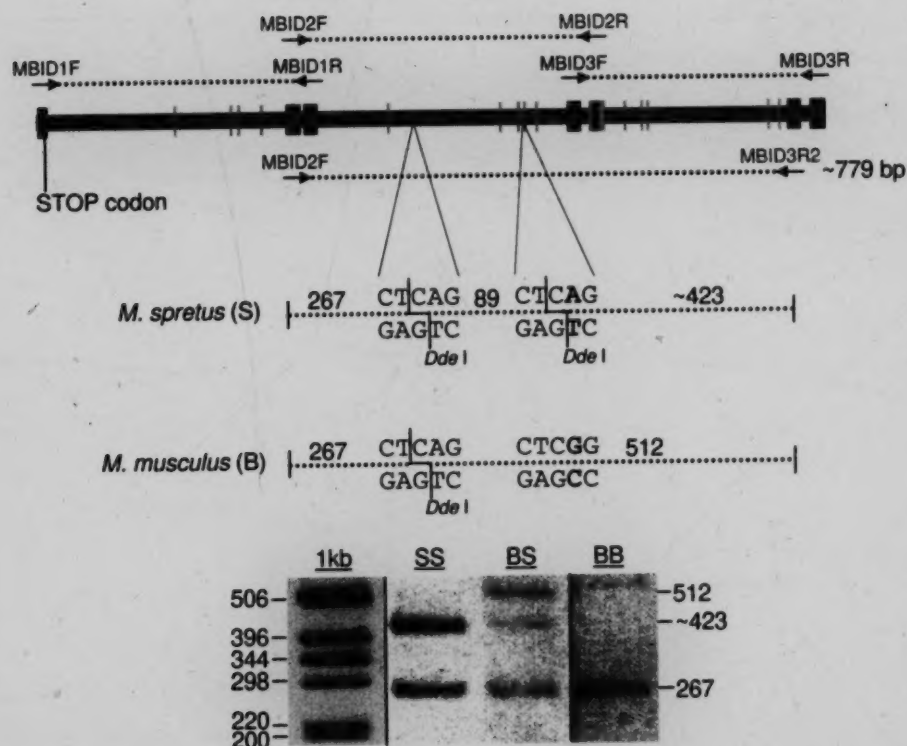


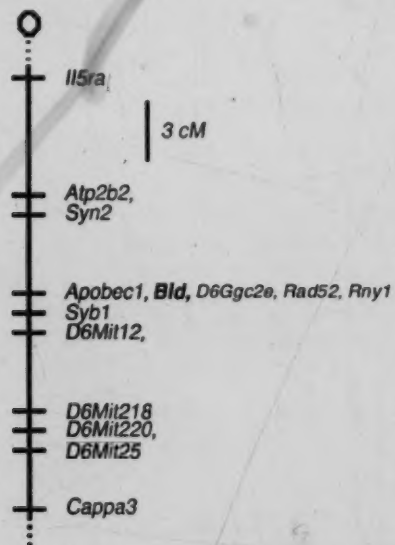
Figure III-7. Sequence differences in the 3' UTRs of *Mus musculus* and *M. spretus Bid*. The three PCR products indicated above the bar were partially sequenced and identified 21 base substitutions and one insertion/deletion, at the sites in pink. Typing of the Jax BSS backcross panel utilized primers MBID2F and MBID3R2 to amplify ~779 bp products from each animal, that were susceptible to *Dde* I digestion at two sites for each *M. spretus* allele, but at only one site for each *M. musculus* allele, giving the restriction patterns show below. The BB restriction pattern is from the homozygous *M. musculus* parent, while the SS (*M. spretus* homozygote) and BS (heterozygote) patterns are samples of N2 progeny DNA. The 89 bp band of *M. spretus* alleles was often too faint for visualization

Figure III-8. Mapping of *Bid* to mouse chromosome 6. (A) Haplotype analysis of part of chromosome 6 showing loci from the Jax BSS backcross linked to *Bid*. Loci are listed in order with the most proximal at the top. The white boxes represent the *M. spretus* allele, and the black boxes represent the *M. musculus* allele. The number of animals sharing each haplotype is shown below each column of boxes. The percentage recombination (R) between adjacent loci is shown at the right. Missing typings were inferred from surrounding data where assignment was unambiguous. **(B) Partial map of chromosome 6 from the Jackson BSS backcross. A 3 cM scale bar is shown.** Loci mapping to the same position are listed in alphabetical order. Raw data from The Jackson Laboratory were obtained from <http://www.jax.org/resources/cmdata>. These panels are slightly modified versions of those provided by Dr. M. Barter (The Jackson Laboratory).

(A) Jackson BSS Chromosome 6

D6Mit10	■	■	■	■	■	■	R
Syn2	■	■	■	■	■	■	1.06
Apobec1, <i>Bid</i> ,	■	■	■	■	■	■	4.25
Syb1	■	■	■	■	■	■	1.06
D6Mit12,	■	■	■	■	■	■	1.06
	40	47	1	1	3	1	1

(B) Jackson BSS Chromosome 6



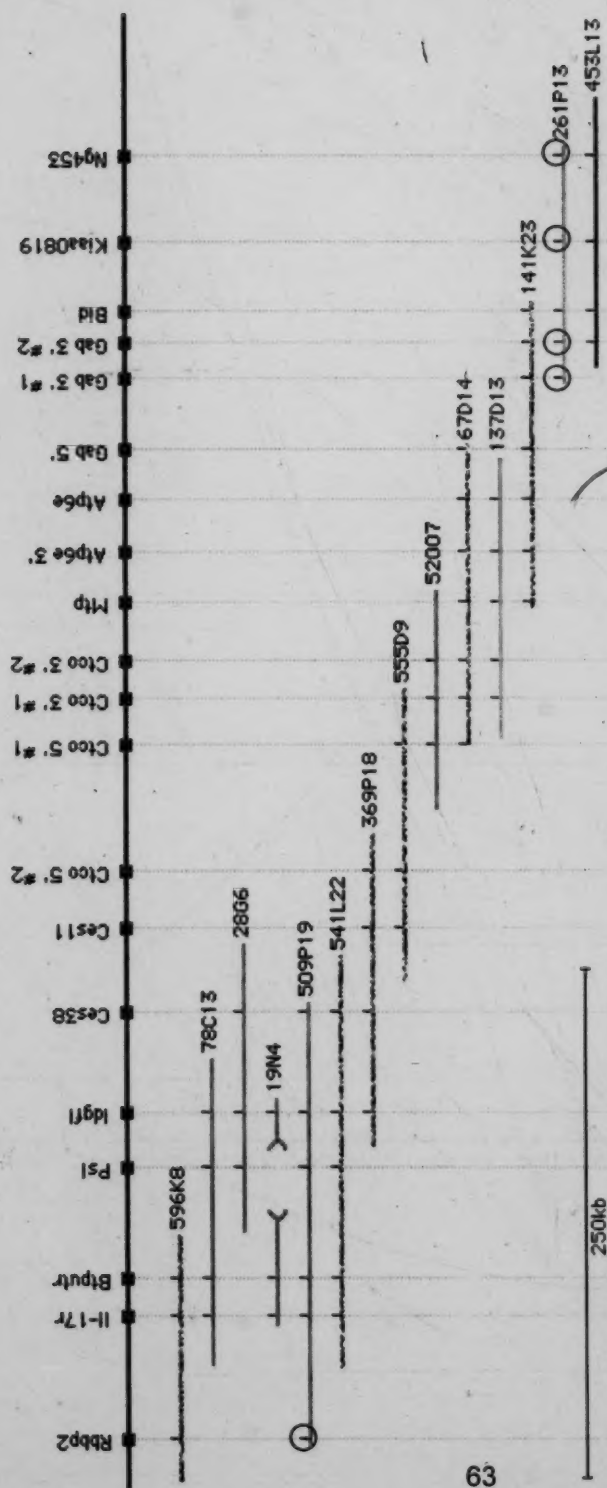


Figure III-9. Physical contig of BAC clones on mouse chromosome 6, in a region homologous to the human CESR on 22q11.2. Stippled bars represent clones which are incompletely sequenced (with BAC 453L13 as the only completed clone as of August 20, 1999; http://www.genome.ou.edu/mus_totals.html) while those that are gray will not be sequenced. All BACs were sized by PFGE and were drawn (with a 250 kb scale bar shown) by ZAPMAP, although the extent of overlapping may not be accurate. The probes shown at the top were mapped by hybridization or by sequence analysis (circles represent analyses that were not performed due to unavailability of sequence) but the distance between each marker is not shown to scale. The orientation of the contig on chromosome 6 is unknown.

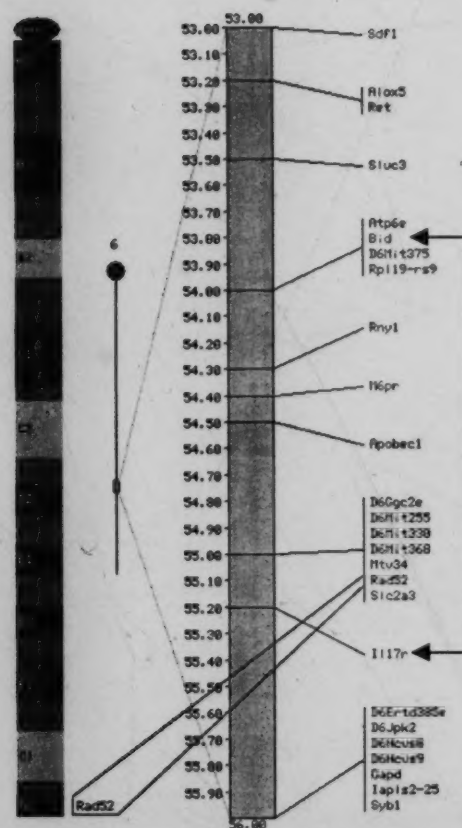


Figure III-10. Partial genetic and cytogenetic maps of mouse chromosome 6. Data depicted in the 53-56 cM interval of the genetic map (right) were obtained from the "MGD stored values" at http://www.informatics.jax.org/searches/linkmap_form.shtml. MGD (Mouse Genome Database) map assignments are assembled from various sources for chromosome committee reports, representing a "consensus map", and thus may contain errors in marker order over short intervals. The close proximity of *Bid* to *Il-17r* is indicated (*Atp6e* is shown adjacent to *Bid*). *Bid* was mapped with the Jax BSS panel, while *Il-17r* was mapped with the Seldin backcross panel (Yao *et al.*, 1995) and *Atp6e* with the Copeland-Jenkins backcross panel (Puech *et al.*, 1997). The cytogenetic mapping of nearby *Rad52* to band G3 (Muris *et al.*, 1994) was illustrated graphically from <http://www.informatics.jax.org/searches/cytomap.cgi>, and a portion of the figure is shown here on the left.

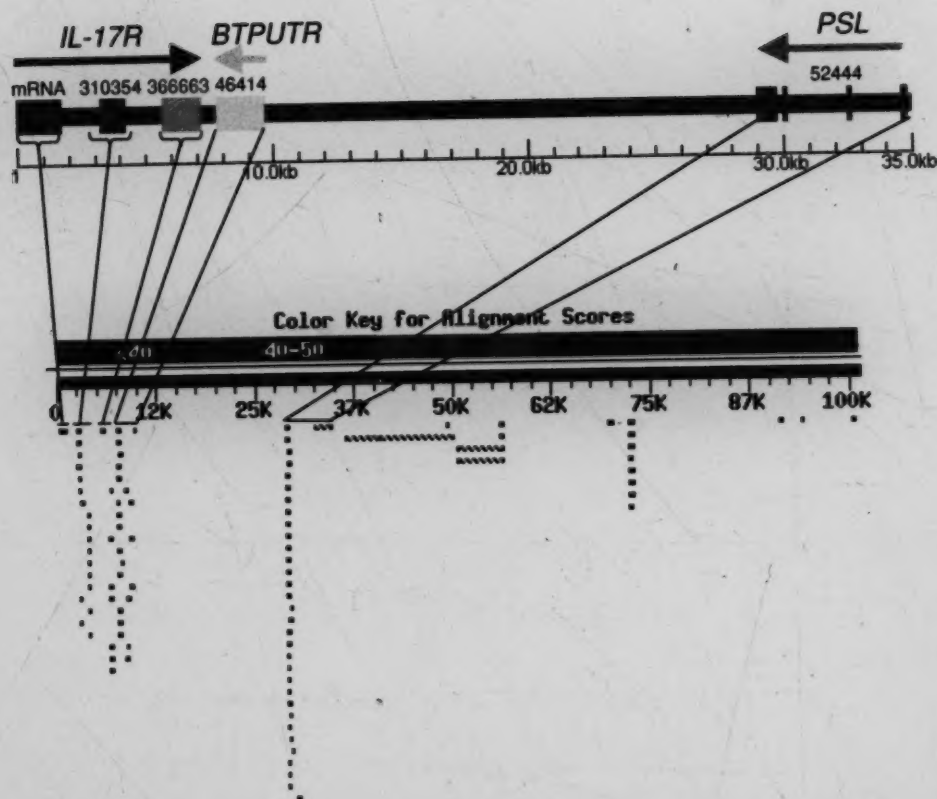


Figure III-11. EST clusters on PAC 143I13. The 101 kb PAC was submitted to a BLASTN search of dbest, after repeat-masking, in centromere-telomere orientation. The graphical output of the search is shown at the bottom, with color-coded "alignment scores" (http://www.ncbi.nlm.nih.gov/BLAST/blast_help.html). Lines are shown to connect the EST clusters to representative cDNA clones discussed in the text.

Figure III-12. Comparative sequence analysis of the genomic regions surrounding human and mouse *GAB/Gab*. Human sequence (top portion) from an area just downstream of *ATP6E* to the distal end of PAC 1087L10 (133 kb in total) was repeat-masked and then compared to assembled contigs of mouse sequence (bottom portion) from BACs 141K23 and 453L13 (the three contig breaks in the partially-sequenced BAC 141K23 are shown as dotted double-lines). Homologous regions identified by "BLAST 2 Sequences" are shown as pink bars, labeled with lower-case letters and connected by lines. Green blocks represent exons of *GAB/Gab* determined from cDNA sequences and are numbered 1-8 for the human gene and 1-7 for the mouse gene. *BID/Bid* exons (1-6) are shown as yellow blocks (as determined from sequenced cDNAs and the published mRNAs). *BID* exon 1 is not located on PAC 1087L10. The area where the longest *GAB* and *BID* transcripts overlap is shown in dark blue. Diagrams above and below the figure illustrate the proposed structures of the transcripts identified by Northern and EST analysis, as well as the human cDNA inserts hybridized to the Northern blots (see text and Figures III-4 and III-13). A 2.4 kb transcript for *GAB* could not be constructed from the known exons and putative polyadenylation signals (shown as light blue bars). Translation initiation codons (for methionine) are depicted as orange bars while translation "stop" codons are red.

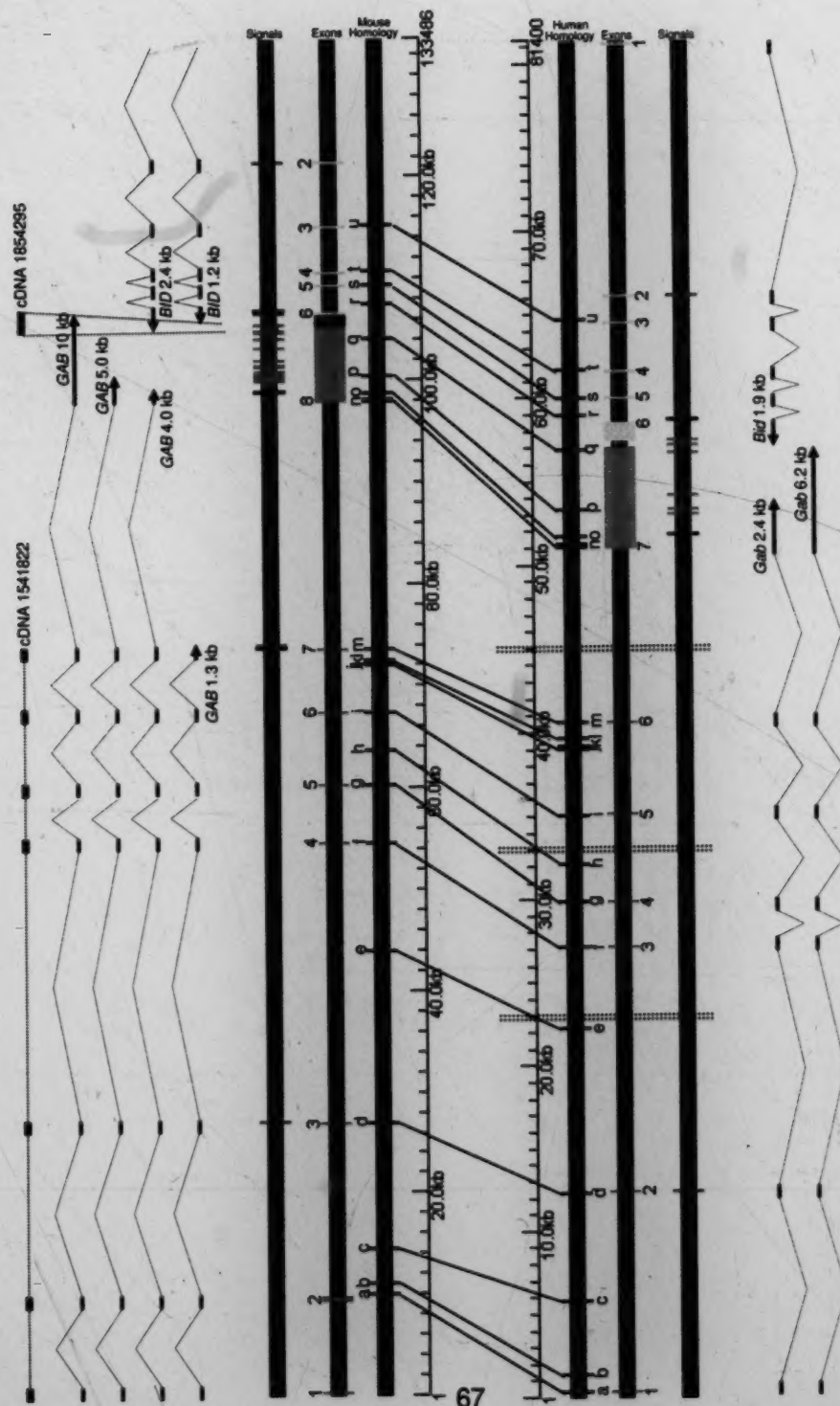


Figure III-13. Northern analysis of *GAB*. Commercial human adult and fetal multiple tissue blots (Clontech, Inc.) were hybridized with radiolabeled human cDNA clone 1541822 (top panels) and subsequently, after removal of bound probe, with human cDNA clone 1854295 (middle panels). Removal of bound probe was performed according to the instructions supplied with the STRIP-EZ DNA kit (Ambion). Post-hybridization washes for 1541822 were 20 min at low stringency (25°C), followed by exposure at -70°C for two days (further washing in high stringency at 50°C only decreased the intensity of each band). Blots probed with 1854295 were washed for 20 min at low stringency (25°C), 20 min at high stringency (10 min each at 50°C and 55°C) and was followed by exposure at -70°C for three days. The bottom panels provide a comparison to the signal from *GAPD* for normalization.

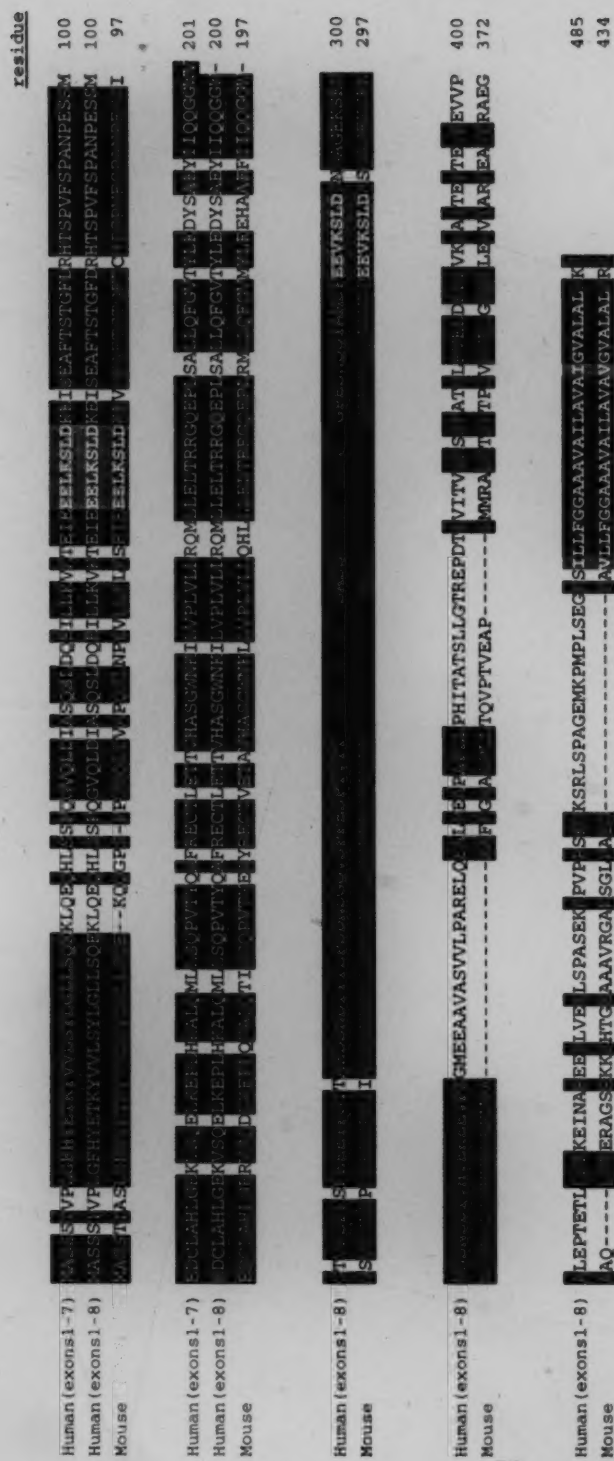


Figure III-14. Amino acid alignment of putative human and mouse GAB proteins. Identical residues are shaded in black. The repeated EE(L/V)KSLD elements are shown in green. The repeated SW(QS/HT)ESLPVSL elements are shaded red. Putative transmembrane domains (as predicted by HMMTOP) are shaded blue. The truncated 201 amino acid human isoform contains only the first element.

Figure III-15. Comparative sequence analysis of the genomic regions surrounding human and mouse *ATP6E/Atp6e*. 36 kb of human sequence (top portion) from clones 1087L10, 273A17 and 77H2 was assembled into three contigs and repeat-masked. It was then compared to two assembled contigs totaling 29 kb from mouse BAC 141K23. Contig breaks are shown as dotted double-lines. "BLAST 2 Sequences" identified the homologous regions ("a" to "m") shown in pink. Green blocks represent exons 1-9, as determined from the published human and mouse mRNAs. The processed mRNAs are depicted at the very top and bottom of the figure. Orange = translation initiation codon for methionine. Red = translation stop codon. Light blue = polyadenylation signals. Purple = Grail-predicted CpG islands. Yellow = NNPP-predicted promoter.

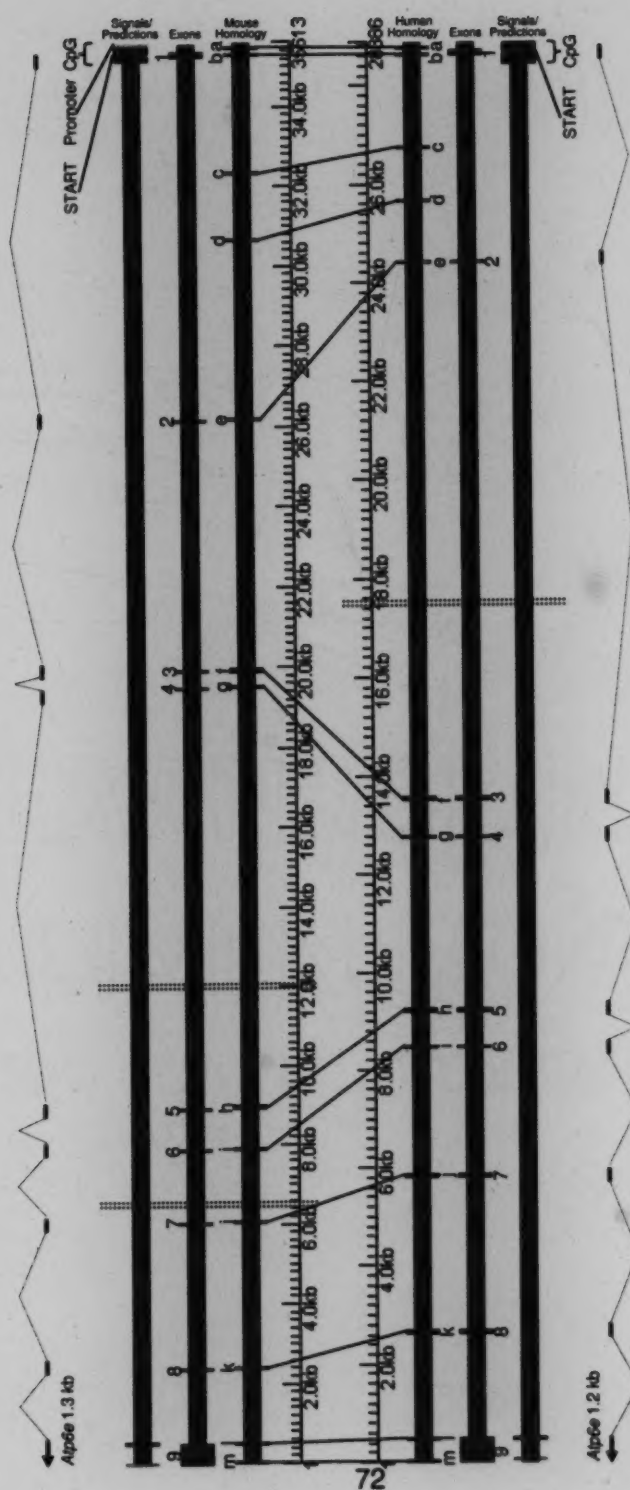


Figure III-16. Comparative sequence analysis of the genomic regions surrounding human and mouse *MTP/Mtp*. 33 kb of human sequence (top portion) from BAC 77H2 was repeat-masked and compared to two assembled contigs totaling 25 kb from mouse BAC 67D14. The contig break is shown as a dotted double-line. "BLAST 2 Sequences" identified the homologous regions ("a" to "j") shown in pink. Green blocks represent exons 1-11 for the human gene, as determined from cDNA sequence (G. Banting, unpublished). Mouse exons 1-10 are also shown in green. Exons 1-7 and most of exons 8 and 10 were determined from the sequence of mouse cDNA clone 404173, while exon 9 (and the whole of 8 and 10) were predicted by GENSCAN and homology to the corresponding human exons. The processed mRNAs are depicted at the top and bottom of the figure. Orange = translation initiation codon for methionine. Red = translation stop codon. Dark blue = GENSCAN-predicted exons. Light blue = polyadenylation signals. Yellow = NNPP-predicted promoters.

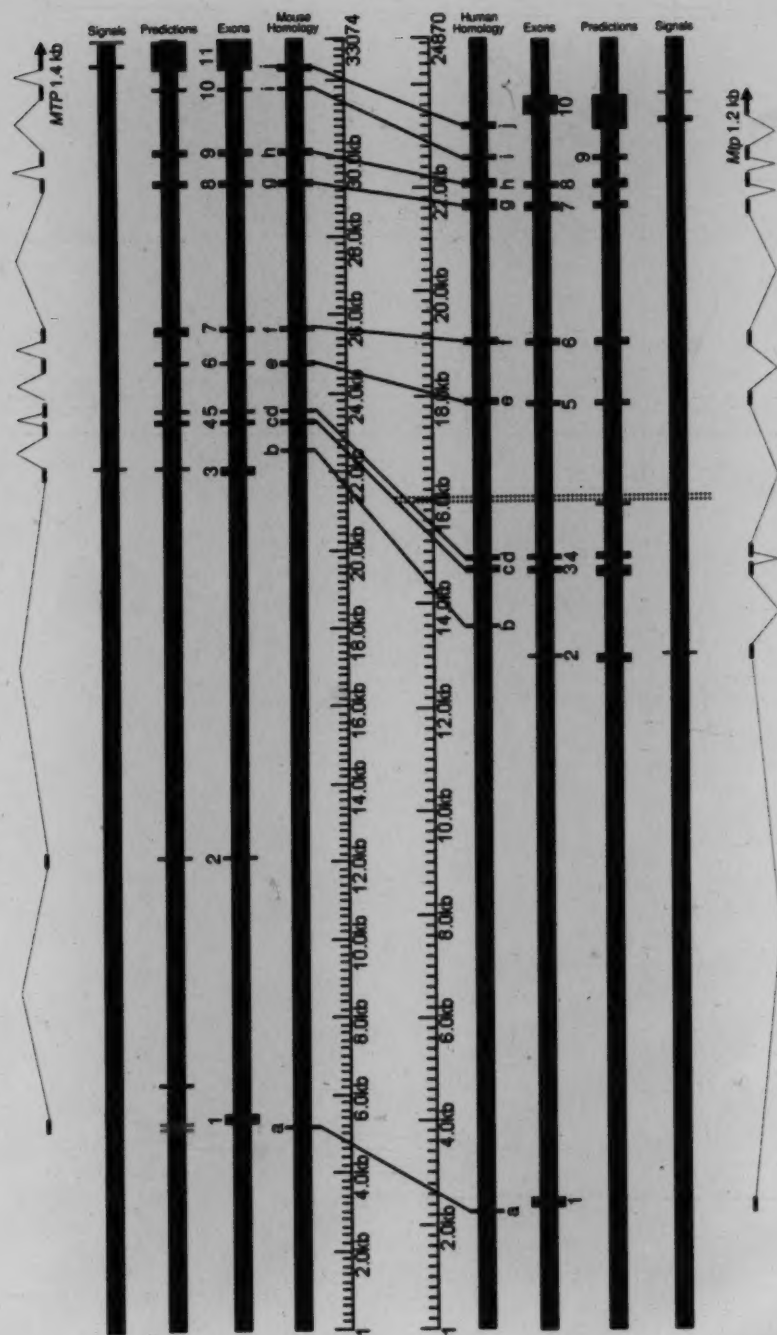


Figure III-17. Prediction of the genomic organization of *BTPUTR*. A 6 kb region of human PAC 143I13 is illustrated, with the location of cDNA 46414 shown in green (3' - 5' orientation; the chromosome 22 centromere is to the left of the figure). Putative ORFs >50 amino acids, translated from the "minus" strand, are shown in the middle. They are bounded by stop codons on the left (corresponding to the tip of the arrow) and potential translation initiation codons (for methionine) on the right. Homology to the partially-sequenced mouse region is shown in pink (gaps on this bar represent either the lack of homology or the lack of sequence information). The GENSCAN-predicted ORF is shown in dark blue, with the putative initiation and STOP codons in orange and red, respectively. The proposed single-exon mRNA of 5-6 kb is depicted at the bottom. Purple = Grail-predicted CpG island. Light blue = polyadenylation signals. Yellow = NNPP-predicted promoter.

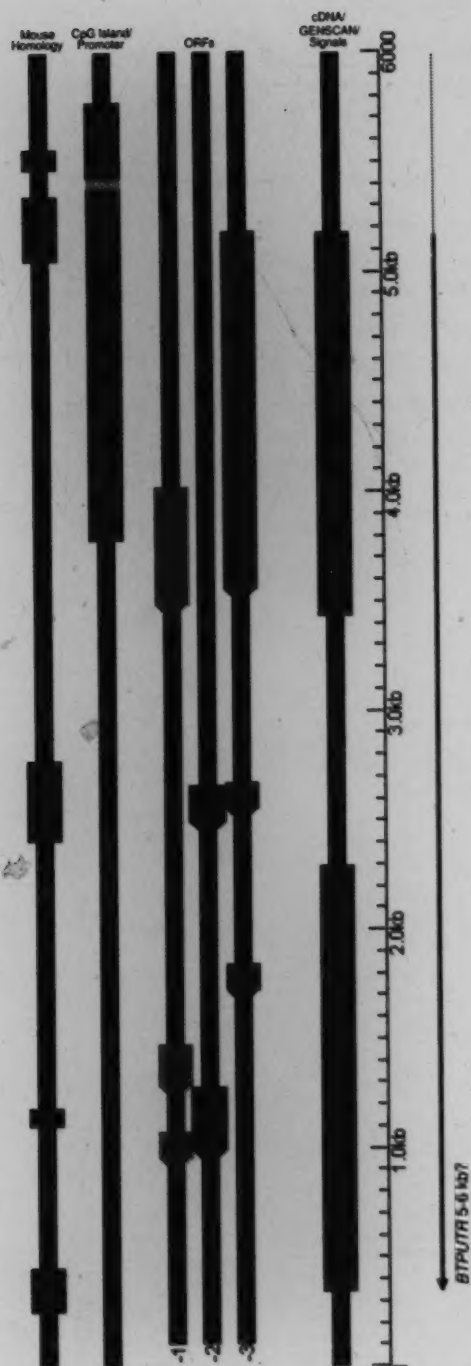


Figure III-18. Organization of the genomic region of *SAHL*. **(A)** A summary of the duplications depicted in Appendix II is shown, clearly indicating those regions that are shared between more than one chromosome. **(B)** In the distal portion of these duplicated segments, the gene *SAHL* appears to have been created from the "shuffling" of fragments with different chromosomal origins. The top two bars reiterate the highest-scoring percent-identities to duplicated genomic regions described in Figure III-18. The third bar illustrates EST hits with the GenBank accession nos., followed by the percent-identity attributed to the highest-scoring segment. The direction of transcription of each cDNA is indicated by arrows (all indicate a 5' - 3' orientation; the chromosome 22 centromere is located to the left of the figure). Green = *SAHL* exons described in Riazi (1998; and unpublished data). Light blue = polyadenylation signals. Purple = Grail-predicted CpG island, adjacent to GENSCAN-predicted exons (in deep blue). Proposed structures of the alternative processed mRNAs are depicted at the bottom.



Chapter IV: Discussion

The CESC-R-homologous region in the mouse demonstrates conserved linkage of ten genes in the interval from Il-17r to Gab

The smallest region duplicated in a cat eye syndrome patient who displayed all of the major CES features (anal atresia, ocular coloboma, congenital heart defect, abnormal kidneys and ear pits and tags) is an ~2 Mb interval from the centromere of human chromosome 22 to between the markers *ATP6E* and *D22S57*, in band 22q11.2 (patient 25105; Mears *et al.*, 1995). The discovery and mapping of *BID* beyond the distal boundary, further delineated the area that contains the candidate genes for producing the CES phenotype. Exon 3 ("70A6-380") of *BID* is not duplicated in patient 25105, as demonstrated by phosphorimaging quantification of the signal emitted from patient DNAs hybridized to a radiolabeled DNA probe. Therefore, *BID* is not predicted to be overexpressed in this patient, and as such, represents the most proximal transcription unit known to be beyond the region critical to the development of CES. Furthermore, the sequence of PAC 1087L10 reveals that marker *D22S57* lies within the first *BID* intron, distal to exon 2 (the gene is transcribed in the orientation of telomere to centromere; Figures IV-1 and III-12). In order to narrow the region to search for candidate genes for CES, the extent of the interstitial duplication of 22q carried by patient S.K. (Knoll *et al.*, 1995) was considered. Marker *D22S795* is not duplicated in this patient (Mears *et al.*, unpublished), therefore, the proximal boundary of his chromosomal rearrangement lies within 1 Mb proximal to *BID* (based on the PFGE map in McDermid *et al.*, 1996). Although S.K. presented with only the heart defect TAPVR, an absent kidney, preauricular pits and mental retardation (in addition to other "minor" CES-related features), examination of the interval from *D22S795* to *BID* (the CESC-R) should allow for the discovery of genes responsible for at least several if not all of the "major" CES features. It is possible that duplication of the genes within this

interval can cause anal atresia and ocular coloboma, as these symptoms do not exhibit complete penetrance when patients with similar supernumerary chromosomes are compared.

Hybridization- and sequenced-based studies in the mouse have demonstrated conservation of linkage of most of the genes between the interval from D22S795 to *BID* and the homologous interval on mouse chromosome 6 (Figure IV-1). After establishing genetic linkage of mouse *Bid* to the distal region of chromosome 6, using the Jax BSS backcross panel, murine cDNA probes were designed from available human and mouse genomic sequence. Partial mouse cDNAs proved to be useful probes for identification of BAC clones, which were assembled into a contig of >600 kb from mouse chromosome 6. This contig spans the region from *Rbbp2* to *Ng453*, encompassing the interval homologous to the region of human chromosome 22 between *IL-17R* and *BID*. As *SAHL* is located distal to marker D22S795, *SAHL* is within the CESCO but outside of the region of conserved linkage with mouse chromosome 6 (see below).

The order of the twelve known genes within and immediately distal to the CESCO, and their homologues in a region of mouse chromosome 6 (the "mCescr"), is conserved. The relative orientation of *CTCO*, *MTP*, *ATP6E*, *GAB* and *BID*, and their mouse homologues, is also conserved, as evidenced from genomic sequence analysis. Preservation of gene orientation in the remainder of the region will be known when complete genomic sequence becomes available for both species. Thus far, relatively few instances in divergence of gene/exon spacing have been demonstrated in the partially-sequenced mouse genes of the mCescr. The overall spacing of homologous DNA elements has been maintained, but is slightly more compact in mouse. This divergence might be attributed to more extensive integration of interspersed repeats and gene duplication in the pericentromeric human region, but confirming this requires more-detailed genomic sequence annotation. In summation, ten CES candidate genes are mapped within the region of human 22q11.2 homologous to mouse chromosome 6: *IL-17R*, *BTPUTR*, *PSL*, *IDGFL*, *CES38*, *CES11*, *CTCO*, *MTP*,

ATP6E and *GAB* (Figure IV-1). Although the localization of mouse *Idgfl* and *Ces11* was based on weakly-hybridizing cross-species probes, their specific hybridization patterns place them in homologous positions of the two physical maps. Their characterization is ongoing (Riazi *et al.*, submitted; P. Brinkmann-Mills, unpublished) and will be aided by full genomic sequence.

It is currently unknown whether any part of *GAB* is duplicated in patient 25105, and thus, whether the gene is entirely or partially within the CESC or not. It is possible that the distal CESC breakpoint lies within the genomic region of *BID*, distal to the 3' end of the longest *GAB* transcript but proximal to *Bid* exon 3 (Figure III-12). The boundary may also lie anywhere downstream of the shortest *GAB* transcript containing exon 8 (with the predicted transmembrane-domain), thus still allowing overexpression of both of the *GAB* gene products in patient 25105. As well, the distal boundary could lie just beyond exon 7 to allow for overexpression of only the short isoform of *GAB*. As phosphorimaging quantification of Northern analysis of *BID* has shown (Figure III-6), it is a relatively simple matter to test what *GAB* messages (if any) are overexpressed in patient 25105. Total RNA from an available lymphoblastoid cell line from this patient could be probed with cDNA 1541822 which binds to all predicted mRNA molecules for *GAB*. Comparison to a normal control should indicate which bands emit more signal in the patient, and therefore which DNA sequences are duplicated in his genome, thereby further delineating the distal CESC boundary. Alternatively, cDNA probes from the region could be used to investigate the nature of the breakpoint on patient 25105's supernumerary (double-ring) chromosome by Southern hybridization to genomic DNA.

The region of conserved synteny corresponding to the CESC and mouse chromosome 6 does not include the DiGeorge syndrome (DGS) critical deletion region. The DGS region starts 1.6 Mb distal to marker D22S57/*BID* on chromosome 22 (based on the PFGE map in McDermid *et al.*, 1996). The location of the DGS-homologous interval is on mouse chromosome 16 (Puech *et al.*, 1997). Therefore the evolutionary rearrangement breakpoint between these two human syndrome-involved linkage groups is located within a <1.6 Mb region

on chromosome 22, between *KIAA0819* (distal to *BID*) and marker D22S36 of the proximal DGS boundary. It is of interest to note that the 250 kb proximal to D22S36 contains a chromosome-22-low-copy-number-repeat (the "LCR22", consisting of at least five genes or pseudogenes) thought to be involved with other LCR22s that participate in unequal sister chromatid exchange events that cause DGS deletions and CES chromosome rearrangements (Edelmann *et al.*, 1999). It is unknown if the proximal LCR22 is associated with the evolutionary rearrangement(s) that resulted in the current pattern of conserved linkages. The genomic regions adjacent to mouse *Il-17r* contain *Btputr* on one side and *Rbbp2* on the other. The location of human *RBBP2* on chromosome 12p13.3 (http://www.hgsc.bcm.tmc.edu/seq_data_old/cgi-bin/bcm-web-regions.cgi?region=12p13.3), suggests that mouse *Rbbp2* is the most proximal marker within a large region that shares conserved synteny with human 12p13 (Figure IV-2). The mCescr would then be oriented proximal to this region, with *Il-17r* closer to the telomere and *Ng453* closer to the centromere. Human homologues of genes immediately proximal of *Ng453* would be located on either 22q11.2, 7, or 10q11.2, depending on whether human *NG453* is on 22q11.2 and whether the compiled genetic mapping data in Figure IV-2 represents the actual relative order of the genes *Sdf1/Alox5/Ret* and *Rny1* (homologous to human 10q11.2 and 7, respectively). The current genetic map is inconsistent with the physical map of the mCescr as size limitations preclude the mCescr from containing the interval *Rny1* - *Slc2a3* illustrated between *Bid* and *Il-17r*. Such inconsistencies can result when genetic maps are determined from compiled data from different sources (see Figure III-10). It is still possible, however, that the mCescr is located within the region of conserved synteny shared with human 12p13, and that *Rbbp2* may be centromeric of *Il-17r*, with the remaining homology to 12p13 distal to *Ng453*.

The 400 kb region centromeric of human *IL-17R* in the CESC is composed of gene fragments with various chromosomal origins. The fact that the closest duplicated region to *IL-17R* is derived from a region of 12p13 that is relatively near to *RBBP2* (~4 cM; http://www.hgsc.bcm.tmc.edu/seq_data_old/cgi-bin/bcm-web-regions.cgi?region=12p13.3) raises the question of whether this

is coincidental, or if the genes currently within 22q11 and 12p13 have been associated until recently in human evolution. The region of paralogy between 22q11 and a genomic clone from 12p13.3 (GenBank accession no. AC005841) exhibits several local alignments of 83-85% identity, but are interrupted by interspersed repeats on both chromosomes (data not shown). An estimate of when the duplication occurred can be calculated from the rate of mutation for intronic sequences of 5×10^{-9} - 7×10^{-9} mutations per site per year (Eichler *et al.*, 1996). The highest-scoring aligned region shows 83% identity over 776 bp, suggesting the duplication happened 24-34 million years ago. An adjacent segment of 22q11 is paralogous with chromosome 16 (GenBank accession no. AC003034), with similar levels of identity and intrusion of repeats. Taken together, and the probability that the entire first Mb of 22q is also composed of gene duplications, these data imply that the association of *IL-17R* and sequences from human 12p13.3 in both man and mouse is probably coincidental. The fact that another region of paralogy with 12p13.3 containing the *von Willebrand factor* pseudogene (position 5-20K; Figure III-18) may mean that 12p13.3 is prone to being duplicated to pericentromeres. Eichler *et al.* (1996) proposed that regions rich in CAGGG motifs may act as hotspots for duplications and for integration of the transposed segment. Only one region particularly rich in CAGGG pentanucleotides (or CCCTG for an element on the opposite strand) could be identified in the proximal CESC (sequences from 12p13.3 were not studied). A 2278 bp segment at position 187.5K (Figure III-18) contains 19 occurrences of CAGGG or CCCTG accounting for 4% of the DNA. These elements occur here more frequently than expected by chance, calculated as 1 per strand per (4^5+4) bp, or 4.4 in a 2278 bp interval (0.2%). These motifs may have been involved in duplication of the proximal region with segments similar to chromosomes 2, X, Y and 19.

The existence of chromosome 22-derived transcripts from the 400 kb duplication "graveyard" implies that accumulation of duplicated gene segments provides opportunities for evolution of novel genes. Known exons of the gene *SAHL* are derived from regions paralogous to parts of three other chromosomes

12, 16 and 21. The high similarity of these regions, including intronic sequence, to other regions of the human genome argues that there is no direct mouse homologue of *SAHL*, as it emerged relatively recently. An interval of 23 kb surrounding the first exon of *SAHL* is 96.2% identical to chromosome 21q11, even across the interspersed repeats. From the estimation method described above, this region was duplicated between 5.4 and 7.6 million years ago, well after the divergence of mouse and man about 80 million years ago. If one assumes that none of the exons of *SAHL* are ancestral to 22q11, then it is extremely unlikely that the mouse genome would have undergone similar genomic shuffling events to create a gene that is functionally similar to *SAHL*. If portions of *SAHL* are deemed to be ancestral to 22q11, then it is unclear why the homologous mouse region (BAC 596K8) does not display any similarity to sequences proximal to *IL-17R*. Based on this evidence, and the extent of gene duplication associated with proximal 22q11, it is likely that none of the genes centromeric of *IL-17R* have orthologues in the mouse. Therefore, cloning of the mCescr probably represents the final stage in determining the area that could be engineered to mimic the genetic defect associated with CES.

Conservation of gene content between the CECR and mouse chromosome 6

The success of modeling CES in the mouse may rely on the ability of duplicated mouse homologues to function in a conserved manner to those duplicated in CES patients. In this study, the protein products of four genes (or partial genes) from the CECR and their mouse homologues were analyzed for the extent of conservation. Such sequence analyses are incapable of determining the actual functional conservation of gene products, or even of domains within the proteins. However, it is a very rapid method to calculate the percent-identity between polypeptides and it allows for predictions of functional conservation based on previously-studied orthologous pairs.

Nearly 2000 human and mouse homologues were studied by Makalowski *et al.* (1996) at the DNA and protein level for the extent of sequence conservation. They found that the amino acid sequences ranged from 36-100% identical, with an average of 85.4%. The presumptive translation product from the *BTPUTR* locus is also precisely 85.4% identical to the putative mouse homologue. These data apply to only the amino-termini as the mouse locus is incompletely-sequenced at this time. *ATP6E* and *MTP* both demonstrate similarity to their homologues at levels above the average (93% and 88%, respectively). The amino acid conservation of *GAB* and *Gab* is significantly lower at 69%, but, as seen in Figure III-14, the two proteins share stretches of up to 74 residues with perfect identity. However, this value applies only to a comparison of mouse *GAB* to the larger human isoform. Mice are not predicted to express a protein similar in length to the 201-residue (smaller) human isoform, therefore during the course of human evolution, a novel function may have been prescribed to a shorter version of *GAB*. One possibility is that the longer form is membrane-bound (i.e. containing the putative membrane-spanning region) whereas the shorter form is free in the cytoplasm; a comparable situation is documented for the protein tyrosine phosphatase epsilon, with the cytosolic form transcribed by an internal promoter (Tanuma *et al.*, 1999). Some of the most divergent homologues still perform similar functions in both man and mouse (e.g. interleukins and interferons and their roles in antimicrobial host defense; Makalowski *et al.*, 1996), yet mutations in certain highly conserved proteins do not faithfully reproduce similar mutant phenotypes. For example, human *MSH2*, involved in mismatch repair and the development of familial nonpolyposis colon cancer, is 95% identical to mouse *Msh2* (Makalowski *et al.*, 1996), yet *Msh2*-deficient mice frequently develop lymphomas instead of colorectal tumors (reviewed in Bedell *et al.*, 1997). On the other hand, mouse models have been developed for human gene defects where the homologue-identity is similar to or below the average. The copper-transporting Menkes disease gene, *ATP7A*, is 84.4% identical to the mouse homologue (Makalowski *et al.*, 1996), which is in turn responsible for the *mottled* mutation. Different *mottled* variants demonstrate

very similar biochemical and overt phenotypes to symptoms of Menkes patients, including intestinal copper accumulation and brittle depigmented hair (reviewed in Bedell *et al.*, 1997). In addition, mice with reduced expression of the *Cftr* gene, for which the human homologue is responsible for cystic fibrosis (CF), closely mimic the pulmonary disease of CF patients (reviewed in Bedell *et al.*, 1997) even though the proteins are only 78.5% identical (Makalowski *et al.*, 1996). The high conservation of *ATP6E* and *Atp6e* and the protein's general "housekeeping" role (in establishing protonmotive forces across intracellular membranes) suggests that the duplicated mouse gene could behave in a fashion analogous to that in CES patients. It is unknown if *Mtp* expression is limited to the brain and liver like its human counterpart, but the conservation of a putative GATA-factor binding site (and a high protein identity of 88%) may point to conservation of function for this gene. As there is little data to determine the roles of *BTPUTR/Btputr* and *GAB/Gab*, it is difficult to speculate on the conservation of protein function between mice and humans, but the high level of conservation within the "repeated" and putative transmembrane domains of *GAB* (Figure III-14) indicates that both species have maintained the purpose of these domains. Transgenic mice carrying a human BAC containing *MTP* have been created, but the mice do not display any noticeable mutant phenotype (H. McDermid, unpublished). This implies that duplication of *MTP* does not contribute to the development of CES in human, which was presumed from its lack of expression in the main affected organs (eyes, heart and kidneys).

Another major consideration when predicting the functional similarity of the *CESCR* and *mCescr* genes is the conservation of gene regulation. Namely, this concerns the abundance, timing and tissue-specificity of the expressed gene products. While measuring these qualities requires labor-intensive procedures in a model system (and can be virtually impossible with human subjects), clues to conservation of gene regulation can be derived from conservation of noncoding DNA elements (Hardison *et al.*, 1997; Clark, 1999). It is conceivable that regulatory regions of the genes have diverged in human and mouse thus precluding their identification by sequence homology. As well, since most

transcription factor binding sites are only a few nucleotides long, short stretches of regional identity (<20 bp) will be overlooked during comparison by "BLAST 2 Sequences". However, *BTPUTR/Btputr*, *MTP/Mtp*, *ATP6E/Atp6e* and *GAB/Gab* were examined for homology at the DNA level. For each gene, the overall conservation of sequence within coding elements versus noticeably-homologous noncoding elements was comparable. Relative to the *BTK/Btk* locus characterized by Oeltjen *et al.* (1997) to show 73% identity for homologous noncoding segments, the values for the *CESCR/mCescr* genes are consistently higher (79.4% - 91.9%). It is therefore promising that these *CESCR* genes are regulated in manners similar to their mouse homologues, and that engineered duplication of the mouse genes would result in cellular consequences analogous to those that must occur during the development of CES. To address the fact that NCHs were not as pervasive in *MTP* and *ATP6E*, it has been proposed that genes involved in "housekeeping" duties are under simple genetic control and therefore do not require large regulatory regions (Hardison *et al.*, 1997). This proposal seems to contradict data concerning the expression of *GAB*. *GAB* is ubiquitously and abundantly expressed, like many housekeeping genes, yet possesses the largest quantity of homologous noncoding sequences of the genes in this study. Interesting hypotheses to account for this are that the timing of *GAB* transcription in early development is tightly regulated, or that the homologous regions contain elements important for specific chromatin organization. The existence of multiple transcript variants for both *GAB* and *Gab* indicates that regulation of this gene may involve mRNAs with different stabilities, secondary structures or localizations. All of the predictions made concerning transcription factor binding sites (for all of the genes including the GATA-X site in *MTP/Mtp*) should be considered with caution, as the consensus recognition sequences are relatively short and can appear in a given genomic region by chance alone. NCHs appearing in transcribed portions, such as the 3' UTRs, of the *CESCR/mCescr* genes may be involved in coordinating mRNA secondary structure or help to specify post-transcription processing. The developmental timing and the role of different-sized transcripts for *Gab* may be tested by *in situ*

hybridization experiments, but should await the precise mapping of *GAB* relative to the distal boundary of the CESC.

The divergence of proximity of the longest 3' UTR of *GAB/Gab* to that of *BID/Bid* suggests that there is no functional significance to the overlap identified in humans. Many gene pairs with overlapping 3' UTRs have been studied (e.g. Spencer *et al.*, 1986, Petrukhin *et al.*, 1998; Shintani *et al.*, 1999), and proposed hybrid RNA molecules regulating messenger stability have been investigated in a few instances (Kimelman and Kirschner, 1989; Hildebrandt and Nellen, 1992). However, the abundance of different-sized 3' UTRs for both *GAB* and *BID* may instead reflect a regional context that is uniquely inefficient for mRNA truncation, and "accidentally" allows for overlap of certain transcripts. The lack of tissue-specificity for any of the transcripts of these genes may suggest they are not coordinately-regulated.

Practical considerations for creating a mouse model of CES

The aim of this project was to determine the feasibility with which a duplication of mouse genes homologous to those in the CESC could be engineered. As noted above, ten genes from the CESC and their mouse homologues are tightly linked within their respective regions and share conserved order and significant gene content. This interval in humans represents ten of the most proximal genes duplicated in patient S.K. who exhibits some of the major CES defects (TAPVR, abnormal kidneys and preauricular pits) and several minor CES defects including typical facial abnormalities (hypertelorism, downslanting palpebral fissures, epicanthal folds, thin upper lip and flat nasal bridge), genital defects, hypotonia and moderate motor and cognitive delay (Knoll *et al.*, 1995). As these ten genes are also duplicated in every CES patient, with the possible exception of *GAE* in patient 25105, and result in overlapping phenotypes (MIM115470) overexpression of their gene products must contribute to S.K.'s phenotype. An engineered duplication of the

ten homologues in the mCescr could mimic at least some of S.K.'s features. The high variability of the CES phenotype indicates that the CES features not observed in S.K. (notably anal atresia and ocular coloboma), which are sometimes not observed in other CES patients, allows for the possibility that the ten mCescr homologues studied here could also mimic these defects when overexpressed in mice. It is likely that engineered mice would also recapitulate the variable penetrance and severity of CES by producing some recombinant mice with a mild (or absent) mutant phenotype when the duplicated allele is introduced into different genetic backgrounds through controlled crosses. The impact of stochastic factors on the variable expression of CES features might also be recognized from phenotypic variability within a mouse strain. The characterization of mice partially trisomic for chromosome 6 that display ocular coloboma and cleft palate could be a promising indication that mCescr genes are responsible for these defects and act in a manner similar to their human homologues when overexpressed. However, these mice were duplicated for a large portion of chromosome 6 (bands B3-G3; see Figure III-10) and lacked the short distal end of chromosome 13 (Cacheiro *et al.*, 1994). Therefore, the mCescr represents only a fraction of the affected genes in these mice. The fact that genes proximal to *IL-17R* (such as *SAHL*) may also contribute to the etiology of CES precludes the ability to test their dosage-sensitivity through engineered duplication in mice, as mice are not expected to have orthologous loci. Therefore, manipulation of mouse pronuclei by injection of human transgenes appears to be the only way to assay the involvement of these genes in CES, although their overexpression may not have any ascertainable effect in another organism.

While chromosome engineering remains the most efficient method of mimicking a CES-like genomic rearrangement, PAC or BAC transgenesis is an attractive way to quickly test a human gene's dosage-sensitivity in a model system. Due to the size limitation of bacterial clone inserts, the clones mapped to the CESCRR will only contain from one to three complete genes as their insert sizes range from only 100-280 kb (Johnson *et al.*, 1999). Therefore, a single

transgenic mouse would not carry a "duplication" of all of the genes sharing conserved linkage with the mCesr, unless the different transgenic strains were carefully intercrossed and monitored for integrity of the transgenes. However, transgenic mice will be extremely important in narrowing-down the candidate genes that actually cause CES-like defects, as it is suspected that duplication of certain CESCR genes may not produce an overt mutant phenotype. In support of this, preliminary studies of mice carrying a human insert in which *MTP* is the only intact gene do not display noticeable cellular defects (H. McDermid, unpublished). There are other explanations for the inability of a human transgene to model a human condition in a mouse background. These include restricted expression of the human transgene (e.g. eight integrated copies of the human *PMP22* transgene results in only ~1.7-times the expression compared to the two innate copies of the mouse gene; Huxley *et al.*, 1996). As well, a sufficiently high copy number could induce homology-dependent co-suppression of the transgene (Garlick *et al.*, 1998), although this may be avoided by using large insert vectors such as BACs or PACs. Apart from problems with the actual dose of transgene expression, a human gene may not be able to interact with the mouse homologues of the gene products that it interacts with in a human cellular environment to cause CES. While this factor could be remedied by the use of mouse transgenes (which could be derived from the chromosome 6 BAC map presented here), they are much more difficult to study as molecular methods to test for integration and insert-integrity would cross-react with the endogenous mouse genes.

Chromosome engineering (such as the method described in Ramirez-Solis *et al.*, 1995 and Figure 1-2) is a viable alternative to transgenic modeling of CES. Although manipulation of ES cell lines is more technically-demanding than pronuclear injection, it would be possible to duplicate *Il-17r*, *Btputr*, *Psl*, *Idgfl*, *Ces38*, *Ces11*, *Ctco*, *Mtp*, *Atp6e* and *Gab* within a single mouse genome. Controlled crosses would also allow for creation of mice with four copies of the mCesr, more closely matching the genetic defect seen in most CES patients. Insertion of the recombinogenic lox P sites between *Rbbp2* and *Il-17r* and

between *Gab* and *Bid*, would allow for Cre-directed recombination to produce an ES line with a genetically-balanced reciprocal deletion/duplication of the mCesr. Mice carrying the duplication would be easily examined for overt CES-like features such as ocular coloboma, facial dysmorphia and anal atresia. Post-mortem studies would also ensue to study internal organ defects such as congenital heart disease and kidney anomalies. A high rate of *in utero* lethality could point to defects for which mouse development has less "tolerance", but post-mortem examinations would identify the affected organs. Creation of mice with the deletion allele of the mCesr may also be informative of the roles of the missing gene products.

Producing a mouse model of CES creates the opportunity to examine gene dosage effects that affect early embryological development. Since the defects seen in CES patients likely originate during the first ten weeks post-conception, the critical overexpressed genes likely play important roles in early embryogenesis. To better understand these processes and the human disorders that result from disruption of developmental pathways, it is necessary to study a model system which can be manipulated for studying gene action and therapeutic intervention, such as the mouse. Modeling CES offers the chance to study a medical condition for which the genetic duplication can be relatively small (~2 Mb in patient 25105; Mears *et al.*, 1995). This is an advantage over the study of Down syndrome, a much more medically-prevalent duplication disorder, that is typically caused by complete trisomy (of chromosome 21). The use of homology-based physical mapping in the mouse has resulted in identifying a region that exhibits conserved linkage with a significant portion of the human genome that is responsible for causing CES. The tight linkage of genes within the mCesr lends the region to be manipulated by chromosomal engineering that should be a reliable means of modeling the events that lead to CES.

Figure IV-1. The physical order of genes within the CESCO and the homologous region (mCescr) in mouse. Gene orientation (with the arrowhead pointing to the 3' end), where indicated, was determined by genomic sequence analysis. Dark boxes represent genes sharing conserved linkage between human chromosome 22q11.2 and mouse chromosome 6. Lightly-shaded boxes represent other genes characterized in this study: the human homologue of *Rbbp2* is on chromosome 12p13.3; respective homologues of human *SAHL* and mouse *Ng453* were not identified. The orientation of the mCescr on chromosome 6 was assumed from data demonstrating that a large region telomeric from *Il-17r* (mapped by genetic and cytogenetic methods; Figures III-10 and IV-2) contains genes whose human homologues are on chromosome 12p13.3, near human *RBBP2*.

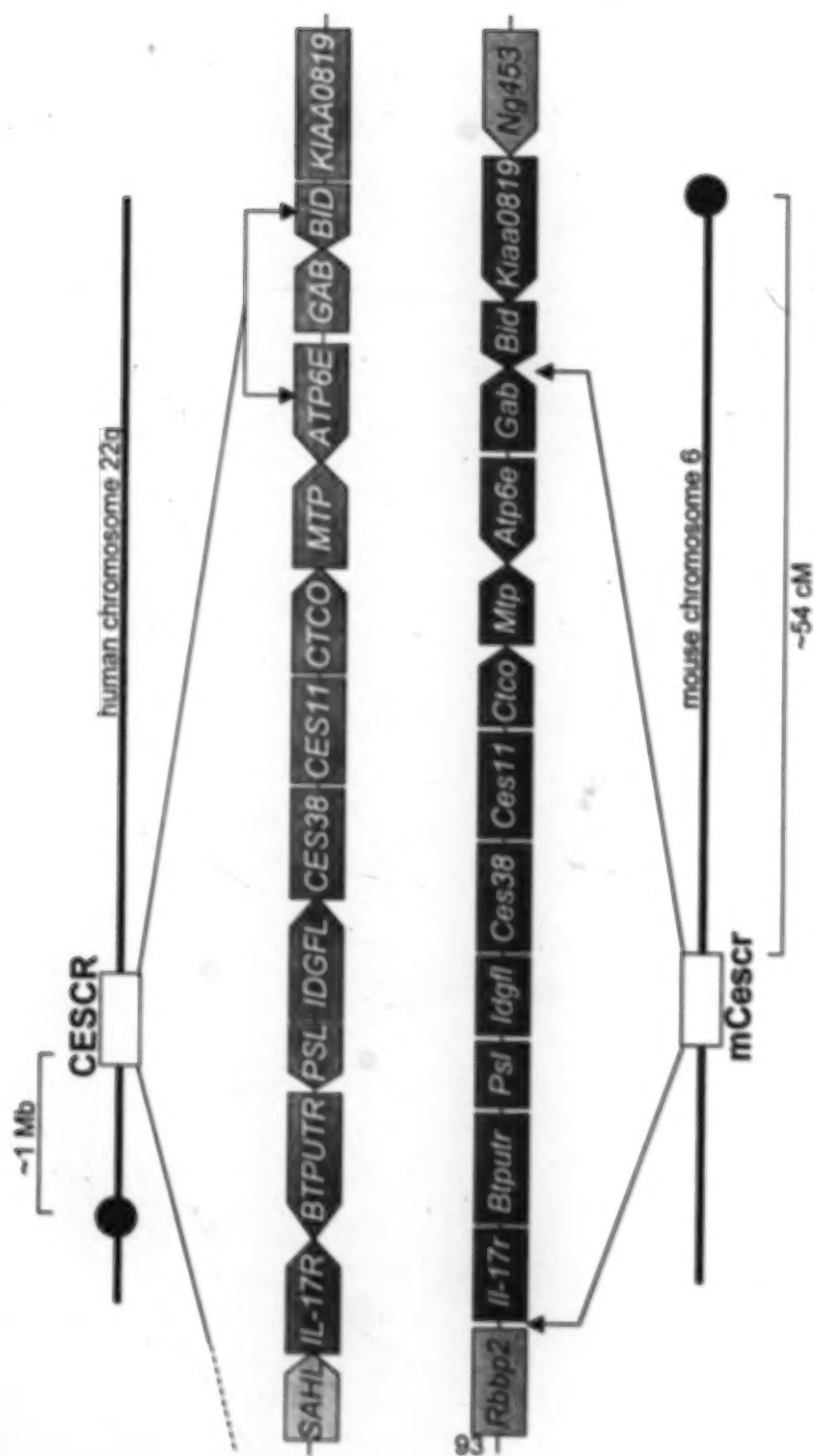
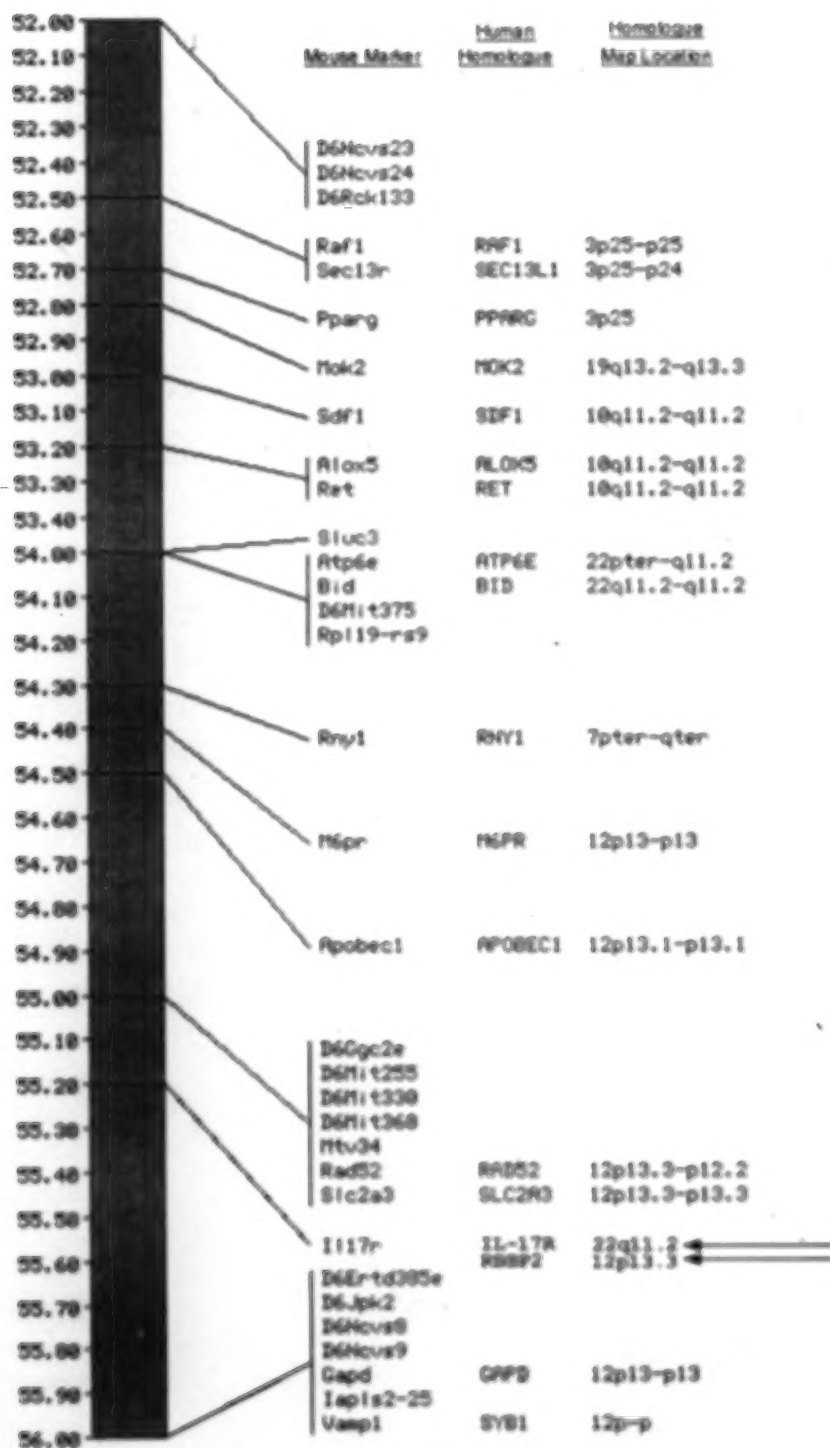


Figure IV-2. The genetic map of mouse chromosome 6 demonstrating the conserved synteny with regions of the human genome. Data depicted in the 52-56 cM interval was obtained from the "MGD stored values" at http://informatics.jax.org/searches/linkmap_form.shtml. The MGD map assignments are assembled from various sources representing a "consensus map", and thus may contain errors in marker order over short intervals. The location of human *IL-17R* on chromosome 22q11.2 was added to the list of "Homologous Map Locations". The *Rbbp2* locus was also added, indicated as being closely-linked to *IL-17r*. The block of conserved linkage shared with 22q11.2 (illustrated in Figures III-9 and IV-1) is flanked by genes whose human homologues map to chromosomes 10q11.2, 7 and 12p13.3.



Bibliography

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* 215: 403-410.
- Ansari-Lari, M. A., Oeltjen, J. C., Schwartz, S., Zhang, Z., Muzny, D. M., Lu, J., Gorrell, J. H., Chinault, A. C., Belmont, J. W., Miller, W., and Gibbs, R. A. (1998). Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res* 8: 29-40.
- Antonarakis, S. E. (1998). 10 years of Genomics, chromosome 21, and Down syndrome. *Genomics* 51: 1-16.
- Ausubel F.M., Brent R., Kingston R.E., Moore D.D., Seidman J.G., A. S.J., and Struhl K. eds. (1989). *Current Protocols in Molecular Biology*. Greene Publishing Associates and Wiley-Interscience.
- Baud, V., Mears, A. J., Lamour, V., Scamps, C., Duncan, A. M., McDermid, H. E., and Lipinski, M. (1994). The E subunit of vacuolar H(+)-ATPase localizes close to the centromere on human chromosome 22. *Hum Mol Genet* 3: 335-339.
- Bedell, M. A., Largaespada, D. A., Jenkins, N. A., and Copeland, N. G. (1997). Mouse models of human disease. Part II: recent progress and future directions. *Genes Dev* 11: 11-43.
- Bell, C. J., Budarf, M. L., Nieuwenhuijsen, B. W., Barnoski, B. L., Buetow, K. H., Campbell, K., Colbert, A. M., Collins, J., Daly, M., Desjardins, P. R., and et al. (1995). Integration of physical, breakpoint and genetic maps of

chromosome 22. Localization of 587 yeast artificial chromosomes with 238 mapped markers. *Hum Mol Genet* 4: 59-69.

Bingham, P. M. (1997). Cosuppression comes to the animals [comment]. *Cell* 90: 385-387.

Bleyl, S., Nelson, L., Odelberg, S. J., Ruttenberg, H. D., Otterud, B., Leppert, M., and Ward, K. (1995). A gene for familial total anomalous pulmonary venous return maps to chromosome 4p13-q12. *Am J Hum Genet* 56: 408-415.

Brendel, V., Bucher, P., Nourbakhsh, I. R., Blaisdell, B. E., and Karlin, S. (1992). Methods and algorithms for statistical analysis of protein sequences. *Proc Natl Acad Sci U S A* 89: 2002-2006.

Brinkman-Mills, P. (1999). Transcriptional mapping in human chromosome 22q11.2. M. Sc. Thesis. University of Alberta.

Buckler, A. J., Chang, D. D., Graw, S. L., Brook, J. D., Haber, D. A., Sharp, P. A., and Housman, D. E. (1991). Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. *Proc Natl Acad Sci U S A* 88: 4005-4009.

Budarf, M. L., and Emanuel, B. S. (1997). Progress in the autosomal segmental aneusomy syndromes (SASs): single or multi-locus disorders? *Hum Mol Genet* 6: 1657-1665.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268: 78-94.

Burn, T. C., Connors, T. D., Klinger, K. W., and Landes, G. M. (1995). Increased

exon-trapping efficiency through modifications to the pSPL3 splicing vector. *Gene* 161: 183-187.

Cacheiro, N. L., Rutledge, J. C., Cain, K. T., Cornett, C. V., and Generoso, W. M. (1994). Cytogenetic analysis of malformed mouse fetuses derived from balanced translocation heterozygotes. *Cytogenet Cell Genet* 66: 139-148.

Carver, E. A., and Stubbs, L. (1997). Zooming in on the human-mouse comparative map: genome conservation re-examined on a high-resolution scale. *Genome Res* 7: 1123-1137.

Clark, M. S. (1999). Comparative genomics: the key to understanding the Human Genome Project. *Bioessays* 21: 121-130.

Claverie, J. M. (1997). Computational methods for the identification of genes in vertebrate genomic sequences. *Hum Mol Genet* 6: 1735-1744.

Collins, J. E., Cole, C. G., Smink, L. J., Garrett, C. L., Levensha, M. A., Soderlund, C. A., Maslen, G. L., Everett, L. A., Rice, K. M., Coffey, A. J., and et al. (1995). A high-density YAC contig map of human chromosome 22. *Nature* 377: 367-379.

Edelmann, L., Pandita, R. K., Spiteri, E., Funke, B., Goldberg, R., Palanisamy, N., Chaganti, R. S., Magenis, E., Shprintzen, R. J., and Morrow, B. E. (1999). A common molecular basis for rearrangement disorders on chromosome 22q11. *Hum Mol Genet* 8: 1157-1167.

Eichler, E. E. (1999). Repetitive conundrums of centromere structure and function. *Hum Mol Genet* 8: 151-155.

Eichler, E. E., Lu, F., Shen, Y., Antonacci, R., Jurecic, V., Doggett, N. A., Moyzis,

- R. K., Baldini, A., Gibbs, R. A., and Nelson, D. L. (1996). Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum Mol Genet* 5: 899-912.
- Erickson, R. P. (1996). Mouse models of human genetic disease: which mouse is more like a man? *Bioessays* 18: 993-998.
- Frangiskakis, J. M., Ewart, A. K., Morris, C. A., Mervis, C. B., Bertrand, J., Robinson, B. F., Klein, B. P., Ensing, G. J., Everett, L. A., Green, E. D., Proschel, C., Gutowski, N. J., Noble, M., Atkinson, D. L., Odelberg, S. J. and Keating, M. T. (1996). LIM-kinase1 hemizyosity implicated in impaired visuospatial constructive cognition. *Cell* 86: 59-69.
- Franklin, R. C., and Parslow, M. I. (1972). The cat-eye syndrome. Review and two further cases occurring in female siblings with normal chromosomes. *Acta Paediatr Scand* 61: 581-586.
- Galili, N., Baldwin, H. S., Lund, J., Reeves, R., Gong, W., Wang, Z., Roe, B. A., Emanuel, B. S., Nayak, S., Mickanin, C., Budarf, M. I., and Buck, C. A. (1997). A region of mouse chromosome 16 is syntenic to the DiGeorge, velocardiofacial syndrome minimal critical region. *Genome Res* 7: 399.
- Garrick, D., Fiering, S., Martin, D. I., and Whitelaw, E. (1998). Repeat-induced gene silencing in mammals [see comments]. *Nat Genet* 18: 56-59.
- Hardison, R. C., Oeltjen, J., and Miller, W. (1997). Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res* 7: 959-966.
- Hildebrandt, M., and Nellen, W. (1992). Differential antisense transcription from

the Dictyostelium EB4 gene locus: implications on antisense-mediated regulation of mRNA stability. *Cell* 69: 197-204.

Huxley, C., Passage, E., Manson, A., Putzu, G., Figarella-Branger, D., Pellissier, J. F., and Fontes, M. (1996). Construction of a mouse model of Charcot-Marie-Tooth disease type 1A by pronuclear injection of human YAC DNA. *Hum Mol Genet* 5: 563-569.

Jang, W., Hua, A., Spilson, S. V., Miller, W., Roe, B. A., and Meisler, M. H. (1999). Comparative sequence of human and mouse BAC clones from the mnd2 region of chromosome 2p13. *Genome Res* 9: 53-61.

Johnson, A., Minoshima, S., Asakawa, S., Shimizu, N., Shizuya, H., Roe, B. A., and McDermid, H. E. (1999). A 1.5-Mb contig within the cat eye syndrome critical region at human chromosome 22q11.2. *Genomics* 57: 306-309.

Keeling, D. J., Herslof, M., Ryberg, B., Sjogren, S., and Solvell, L. (1997). Vacuolar H(+)-ATPases. Targets for drug discovery? *Ann N Y Acad Sci* 1997 Nov 3;834:600-8 834: 600-608.

Kimelman, D., and Kirschner, M. W. (1989). An antisense mRNA directs the covalent modification of the transcript encoding fibroblast growth factor in *Xenopus* oocytes. *Cell* 59: 687-696.

Kishino, T., Lalande, M., and Wagstaff, J. (1997). UBE3A/E6-AP mutations cause Angelman syndrome [published erratum appears in *Nat Genet* 1997 Apr;15(4):411]. *Nat Genet* 15: 70-73.

Knoll, J. H., Asamoah, A., Pletcher, B. A., and Wagstaff, J. (1995). Interstitial duplication of proximal 22q: phenotypic overlap with cat eye syndrome. *Am J Med Genet* 55: 221-224.

- Kohlhase, J., Wischermann, A., Reichenbach, H., Froster, U., and Engel, W. (1998). Mutations in the SALL1 putative transcription factor gene cause Townes- Brocks syndrome. *Nat Genet* 18: 81-83.
- Kola, I., and Hertzog, P. J. (1997). Animal models in the study of the biological function of genes on human chromosome 21 and their role in the pathophysiology of Down syndrome. *Hum Mol Genet* 6: 1713-1727.
- Kozak, M. (1987). At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J Mol Biol* 196:947-950.
- Krantz, I. D., Smith, R., Colliton, R. P., Tinkel, H., Zackai, E. H., Piccoli, D. A., Goldmuntz, E., and Spinner, N. B. (1999). Jagged1 mutations in patients ascertained with isolated congenital heart defects. *Am J Med Genet* 84: 56-60.
- Kuslich, C. D., Kobori, J. A., Mohapatra, G., Gregorio-King, C., and Donlon, T. A. (1999). Prader-Willi syndrome is caused by disruption of the SNRPN gene. *Am J Hum Genet* 64: 70-76.
- Li, D. Y., Toland, A. E., Boak, B. B., Atkinson, D. L., Ensing, G. J., Morris, C. A. and Keating, M. T. (1997). Elastin point mutations cause an obstructive vascular disease, supravalvular aortic stenosis. *Hum Mol Genet* 6: 1021-1028.
- Lindsay, E. A., Shaffer, L. G., Carrozzo, R., Greenberg, F., and Baldini, A. (1995). De novo tandem duplication of chromosome segment 22q11-q12: clinical, cytogenetic, and molecular characterization. *Am J Med Genet* 56: 296-299.

- Lo Nigro, C., Chong, C. S., Smith, A. C., Dobyns, W. B., Carrozzo, R., and Ledbetter, D. H. (1997). Point mutations and an intragenic deletion in LIS1, the lissencephaly causative gene in isolated lissencephaly sequence and Miller-Dieker syndrome. *Hum Mol Genet* 6: 157-164.
- Makalowski, W., Zhang, J., and Boguski, M. S. (1996). Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res* 6: 846-857.
- Mancuso, D. J., Tuley, E. A., Westfield, L. A., Lester-Mancuso, T. L., Le Beasu, M. M., Sorace, J. M. and Sadler, J. E. (1991). Human von Willebrand factor gene and pseudogene: structural analysis and differentiation by polymerase chain reaction. *Biochemistry* 30: 253-269.
- Manson, A. L., Trezise, A. E., MacVinish, L. J., Kasschau, K. D., Birchall, N., Episkopou, V., Vassaux, G., Evans, M. J., Colledge, W. H., Cuthbert, A. W., and Huxley, C. (1997). Complementation of null CF mice with a human CFTR YAC transgene. *Embo J* 16: 4238-4249.
- Matsuura, T., Sutcliffe, J. S., Fang, P., Galjaard, R. J., Jiang, Y. H., Benton, C. S., Rommens, J. M., and Beaudet, A. L. (1997). De novo truncating mutations in E6-AP ubiquitin-protein ligase gene (UBE3A) in Angelman syndrome. *Nat Genet* 15: 74-77.
- McDermid, H. E., Duncan, A. M., Brasch, K. R., Holden, J. J., Magenis, E., Sheehy, R., Burn, J., Kardon, N., Noel, B., Schinzel, A., and et al. (1986). Characterization of the supernumerary chromosome in cat eye syndrome. *Science* 232: 646-648.
- McDermid, H. E., McTaggart, K. E., Riaz, M. A., Hudson, T. J., Budarf, M. L., Emanuel, B. S., and Bell, C. J. (1996). Long-range mapping and

construction of a YAC contig within the cat eye syndrome critical region.
Genome Res 6: 1149-1159.

McKusick, V. A. (1997). Genomics: structural and functional studies of genomes.
Genomics 45: 244-249.

McTaggart, K. E., Budarf, M. L., Driscoll, D. A., Emanuel, B. S., Ferreira, P., and
McDermid, H. E. (1998). Cat eye syndrome chromosome breakpoint
clustering: identification of two intervals also associated with 22q11
deletion syndrome breakpoints. *Cytogenet Cell Genet* 81: 222-228.

Mears, A. J., Duncan, A. M., Budarf, M. L., Emanuel, B. S., Sellinger, B., Siegel-
Bartelt, J., Greenberg, C. R., and McDermid, H. E. (1994). Molecular
characterization of the marker chromosome associated with cat eye
syndrome. *Am J Hum Genet* 55: 134-142.

Mears, A. J., el-Shanti, H., Murray, J. C., McDermid, H. E., and Patil, S. R.
(1995). Minute supernumerary ring chromosome 22 associated with cat
eye syndrome: further delineation of the critical region. *Am J Hum Genet*
57: 667-673.

Muris D. F., Bezzubova O., Buerstedde J. M., Vreeken K., Balajee A. S.,
Osgood C. J., Troelstra C., Hoeijmakers J. H., Ostermann K., Schmidt H.,
et al. (1994). Cloning of human and mouse genes homologous to RAD52,
a yeast gene involved in DNA repair and recombination. *Mutat Res*
315:295-305.

Oakey, R. J., Watson, M. L., and Seldin, M. F. (1992). Construction of a physical
map on mouse and human chromosome 1: comparison of 13 Mb of
mouse and 11 Mb of human DNA. *Hum Mol Genet* 1: 613-620.

Oeltjen, J. C., Malley, T. M., Muzny, D. M., Miller, W., Gibbs, R. A., and Belmont, J. W. (1997). Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res* 7: 315-329.

Patel, P. I., Roa, B. B., Welcher, A. A., Schoener-Scott, R., Trask, B. J., Pentao, L., Snipes, G. J., Garcia, C. A., Francke, U., Shooter, E. M., and et al. (1992). The gene for the peripheral myelin protein PMP-22 is a candidate for Charcot-Marie-Tooth disease type 1A. *Nat Genet* 1: 159-165.

Peterson, K. R., Clegg, C. H., Li, Q., and Stamatoyannopoulos, G. (1997). Production of transgenic mice with yeast artificial chromosomes. *Trends Genet* 13: 61-66.

Petrukhin, K., Koisti, M. J., Bakall, B., Li, W., Xie, G., Marknell, T., Sandgren, O., Forsman, K., Holmgren, G., Andreasson, S., Vujic, M., Bergen, A. A., McGarty-Dugan, V., Figueroa, D., Austin, C. P., Metzker, M. L., Caskey, C. T., and Wadelius, C. (1998). Identification of the gene responsible for Best macular dystrophy. *Nat Genet* 19: 241-247.

Puech, A., Saint-Jore, B., Funke, B., Gilbert, D. J., Sirotkin, H., Copeland, N. G., Jenkins, N. A., Kucherlapati, R., Morrow, B., and Skoultschi, A. I. (1997). Comparative mapping of the human 22q11 chromosomal region and the orthologous region in mice reveals complex changes in gene organization. *Proc Natl Acad Sci U S A* 94: 14608-14613.

Ramírez-Solis, R., Liu, P., and Bradley, A. (1995). Chromosome engineering in mice. *Nature* 378: 720-724.

Rastan, S., and Beeley, L. J. (1997). Functional genomics: going forwards from the databases. *Curr Opin Genet Dev* 7: 777-783.

- Regnier, V., Meddeb, M., Lecointre, G., Richard, F., Duverger, A., Nguyen, V. C., Dutrillaux, B., Bernheim, A., and Danglot, G. (1997). Emergence and scattering of multiple neurofibromatosis (NF1)-related sequences during hominoid evolution suggest a process of pericentromeric interchromosomal transposition. *Hum Mol Genet* 6: 9-16.
- Reiss, J. A., Weleber, R. G., Brown, M. G., Bangs, C. D., Lovrien, E. W., and Magenis, R. E. (1985). Tandem duplication of proximal 22q: a cause of cat-eye syndrome. *Am J Med Genet* 20: 165-171.
- Riazi, M. A. (1998). Transcriptional mapping in the proximal region of human chromosome 22. Ph. D. Thesis. University of Alberta.
- Roa, B. B., Garcia, C. A., Suter, U., Kulpa, D. A., Wise, C. A., Mueller, J., Welcher, A. A., Snipes, G. J., Shooter, E. M., Patel, P. I., and et al. (1993). Charcot-Marie-Tooth disease type 1A. Association with a spontaneous point mutation in the PMP22 gene. *N Engl J Med* 329: 96-101.
- Rossant, J., and Nagy, A. (1995). Genome engineering: the new mouse genetics. *Nat Med* 1: 592-594.
- Rowe, L. B., Nadeau, J. H., Turner, R., Frankel, W. N., Letts, V. A., Eppig, J. T., Ko, M. S., Thurston, S. J., and Birkenmeier, E. H. (1994). Maps from two interspecific backcross DNA panels available as a community genetic mapping resource [published erratum appears in *Mamm Genome* 1994 Jul;5(7):463]. *Mamm Genome* 5: 253-274.
- Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989) *Molecular Cloning, A Laboratory Manual*. Second edition. Cold Spring Harbor Laboratory Press.

- Schinzel, A., Schmid, W., Auf der Maur, P., Moser, H., Degenhardt, K. H., Geisler, M., and Grubisic, A. (1981a). Incomplete trisomy 22. I. Familial 11/22 translocation with 3:1 meiotic disjunction. Delineation of a common clinical picture and report of nine new cases from six families. *Hum Genet* 56: 249-262.
- Schinzel, A., Schmid, W., Fraccaro, M., Tiepolo, L., Zuffardi, O., Opitz, J. M., Lindsten, J., Zetterqvist, P., Enell, H., Baccichetti, C., Tenconi, R., and Pagon, R. A. (1981b). The "cat eye syndrome": dicentric small marker chromosome probably derived from a no.22 (tetrasomy 22pter to q11) associated with a characteristic phenotype. Report of 11 patients and delineation of the clinical picture. *Hum Genet* 57: 148-158.
- Schmickel, R. D. (1986). Contiguous gene syndromes: a component of recognizable syndromes. *J Pediatr* 109: 231-241.
- Shintani, S., O'Huigin, C., Toyosawa, S., Michalova, V., and Klein, J. (1999). Origin of gene overlap: the case of TCP1 and ACAT2. *Genetics* 152: 743-754.
- Smith, A. J., De Sousa, M. A., Kwabi-Addo, B., Heppell-Parton, A., Impey, H., and Rabbitts, P. (1995). A site-directed chromosomal translocation induced in embryonic stem cells by Cre-loxP recombination [published erratum appears in *Nat Genet* 1996 Jan;12(1):110]. *Nat Genet* 9: 376-385.
- Smith, D. J., and Rubin, E. M. (1997). Functional screening and complex traits: human 21q22.2 sequences affecting learning in mice. *Hum Mol Genet* 6: 1729-1733.
- Spencer, C. A., Gietz, R. D., and Hodgetts, R. B. (1986). Overlapping

transcription units in the dopa decarboxylase region of *Drosophila*. *Nature* 322: 279-281.

Stewart, G. D., Harris, P., Galt, J., and Ferguson-Smith, M. A. (1985). Cloned DNA probes regionally mapped to human chromosome 21 and their use in determining the origin of nondisjunction. *Nucleic Acids Res* 13: 4125-4132.

Stewart, G. D., Tanzi, R. E., and Gusella, J. F. (1985). RFLPS at the D21S19 locus of human chromosome 21. *Nucleic Acids Res* 13: 7168.

Suter, U., and Patel, P. I. (1994). Genetic basis of inherited peripheral neuropathies. *Hum Mutat* 3: 95-102.

Tanuma, N., Nakamura, K., and Kikuchi, K. (1999). Distinct promoters control transmembrane and cytosolic protein tyrosine phosphatase epsilon expression during macrophage differentiation. *Eur J Biochem* 259: 46-54.

Tatusova, T. A., and Madden, T. L. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* 174: 247-250.

Tusnády, G. E., and Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol* 283: 489-506.

Vance, J. M., Barker, D., Yamaoka, L. H., Stajich, J. M., Loprest, L., Hung, W. Y., Fischbeck, K., Roses, A. D., and Pericak-Vance, M. A. (1991). Localization of Charcot-Marie-Tooth disease type 1a (CMT1A) to chromosome 17p11.2. *Genomics* 9: 623-628.

- Valentijn, L. J., Baas, F., Wolterman, R. A., Hoogendijk, J. E., van den Bosch, N. H. A., Zorn, I., Gabreels-Festen, A. A. W. M., De Visser, M., and Bolhuis, P. A. (1992). Identical point mutations of *PMP22* in *Trembler-J* mouse and Charcot-Marie-Tooth disease type 1A. *Nat Genet* 2:288-291.
- Wang, K., Yin, X. M., Chao, D. T., Milliman, C. L., and Korsmeyer, S. J. (1996). BID: a novel BH3 domain-only death agonist. *Genes Dev* 10: 2859-2869.
- Xue, Y., Gao, X., Lindsell, C. E., Norton, C. R., Chang, B., Hicks, C., Gendron-Maguire, M., Rand, E. B., Weinmaster, G., and Gridley, T. (1999). Embryonic lethality and vascular defects in mice lacking the Notch ligand Jagged1. *Hum Mol Genet* 8: 723-730.
- Yao, Z., Fanslow, W. C., Seldin, M. F., Rousseau, A. M., Painter, S. L., Comeau, M. R., Cohen, J. I., and Spriggs, M. K. (1995). Herpesvirus Saimiri encodes a new cytokine, IL-17, which binds to a novel cytokine receptor. *Immunity* 3: 811-821.
- Yao Z., Spriggs M. K., Derry J. M., Strockbine L., Park L. S., VandenBos T., Zappone J. D., Painter S. L. and Armitage R. J. (1997). Molecular characterization of the human interleukin (IL)-17 receptor. *Cytokine* 9:794-800 .
- Zhang, M. Q. (1997). Identification of protein coding regions in the human genome by quadratic discriminant analysis [published erratum appears in *Proc Natl Acad Sci U S A* 1997 May 13;94(10):5495]. *Proc Natl Acad Sci S A* 94: 565-568.

Appendix I: GenBank Accession Numbers

Files for the following entries (prominently featured in this thesis) of the GenBank DNA and protein sequence databases can be accessed at <http://www.ncbi.nlm.nih.gov/Entrez/>.

Genes:

Human *BID* mRNA - AF042083
Human BID protein - 2493285
Human *IL-17R* mRNA - HSU58917
Mouse *Il-17r* mRNA - MMU31993
Human *RBBP2* mRNA - NM_005056
Human KIAA0819 mRNA - AB020626

Human ESTs/cDNAs:

128065 - R09650, R09537
1541822 - AA928129
1854295 - AI251761
305358 - W23448
310354 - W30967
366663 - AA026167
46414 - H09224
503784 - AA131711
52055 - H23042, H24321
52444 - H23396
54445 - AA348023, AA348024
548E05 - C17508
589582 - AA146842, AA146843
62073 - AA353887
627125 - AA190546, AA190401

Mouse ESTs/cDNA:

1003369 - AA684166
1003631 - AA683718
1005753 - AA607268
1050354 - AA592135
1181972 - AA711817, AI536427
1382292 - A1462295
303764 - W12281
404173 - W82523, AI425305
464798 - AA030766
516578 - AI429280
556749 - AA104077, AI325043

Human Genomic Clones:

1087L10 - AC006285
109L3 - AC006946
143I13 - AC005300
15J16 - AC005301
20K14 - AC006548
238M15 - AC005399
273A17 - AC007666
357F7 - AC004019
77H2 - AC000052
87O8 - AC007064

283I3 (12p13.3) - AC007406

350L7 (12p13.3) - AC005844

Mouse Genomic Clones:

141K23 - AC006404

369P18 - pending

453L13 - AC006945

541L22 - AC007844

555D9 - pending

596K8 - AC009192

67D14 - AC006447

**Appendix II: Duplicated Segments within the Most Proximal 400
kb Sequenced from the CES Region**

The sequence for this region was assembled from the complete sequences of PACs 15J16, 20K14 and 109L3, and from the five contigs of incompletely-sequenced PAC 87O8 (the gaps are not identified in the figure). Chromosome 22 sequence was repeat-masked and then compared to the nr and htgs databases by BLASTN (in the orientation of centromere-telomere). The graphical outputs were dissected to remove certain low-scoring or redundant hits, then assembled into the annotated diagram. Textual description of the "hits" include the GenBank accession no., followed by the chromosome location (if known) and the percent-identity from the highest-scoring segment. The location of clones sequenced by the Washington University Genome Sequencing Center were discovered by searching with the clone names at http://genome.wustl.edu/gsc/cgi-bin/ace/ctc_choices/ctc.ace.

